Delegates List

Eighth International Strategy Meeting
on Human Genome Sequencing

Genoscope
Evry, France
September 15, 2000

Dr Stephan Beck
The Sanger Centre
The Wellcome Trust Genome Campus
Hinxton
Cambridgeshire CB10 1SA
UK

[redacted]

Dr Ewan Birney
EMBL - EBI
The Wellcome Trust Genome Campus
Hinxton
Cambridgeshire CB10 1SA
UK

[redacted]

Dr Helmut Blöcker
Genome Analysis
GBF
Mascheroder Weg 1
D-38124 Braunschweig
Germany

[redacted]

Dr Allan Bradley
Baylor College of Medicine
One Baylor Plaza
Houston, TX 77030
USA

[redacted]

Dr Francis Collins
National Human Genome Research Institute
National Institutes of Health
31 Center Drive MSC 2152
Building 31, Room 4B09
Bethesda MD 20892-2152
USA

[redacted]

Dr Richard Durbin
The Sanger Centre
The Wellcome Trust Genome Campus
Hinxton
Cambridgeshire CB10 1SA
UK

[redacted]

Dr Nancy Federspiel
Stanford Genome & Technology Center
855 California Avenue
Palo Alto, CA 94304

[redacted]

Dr Adam Felsenfeld
National Human Genome Research Institute
National Institutes of Health
31 Center Drive MSC 2033
Building 31, Room B2B-07
Bethesda, MD 20892-2033
USA

[redacted]

Dr James Harley Gorrell
Baylor College of Medicine HGSC
One Baylor Plaza MSC 226
Houston, TX 77030
USA

[redacted]

Dr Mark Guyer
National Human Genome Research Institute
National Institutes of Health
31 Center Drive MSC 2033
Building 31, Room B2B-07
Bethesda, MD 20892-2033
USA

[redacted]

Dr David Haussler
Computer and Information Science Department
317A Baskin Engineering Building
University of California, Santa Cruz
Santa Cruz, CA 95064
USA

[redacted]

Dr Lee Hood
The Institute for Systems Biology
4225 Roosevelt Way NE, Suite 200
Seattle, WA 98105
USA

[redacted]

Dr Ursula Hurtenbach
DLR – Projekttraeger des BMBF
Gesundheitsforschung
Juedstrasse 125
53175 Bonn
Germany

[redacted]

Dr Rajinder Kaul
University of Washington Genome Center
Box 352145
Seattle, WA 98195
USA

[redacted]

Dr James Kent
Center for Biomolecular Science & Engineering
University of California, Santa Cruz
312B Baskin Engineering
Santa Cruz, CA 95064
USA

[redacted]

Dr Eric Lander
Whitehead Institute/MIT Center for Genomic Research
One Kendall Square
Building 300
Cambridge MA 01239-1516
USA

[redacted]

Dr John McPherson
Washington University Medical School
Genome Sequencing Center
4444 Forest Park Avenue
Box 8501
St Louis MO 63108
USA

[redacted]

Dr Michael Morgan
The Wellcome Trust
183 Euston Road
London NW1 2BE
UK

[redacted]

Dr Donna Muzny
Baylor College of Medicine
One Baylor Plaza
MSC-226
Houston, TX 77030
USA

[redacted]

Dr Richard Myers
Department of Genetics
Stanford University School of Medicine
300 Pasteur Drive, M344
MC 5120
Stanford, CA 94305-5120
USA

[redacted]

Dr Chad Nusbaum
Whitehead Institute for Biomedical Research
MIT Center for Genome Research
320 Charles St.
Cambridge, MA 02141
USA

[redacted]

Dr Ari Patrinos
Biological and Environmental Research
US Department of Energy
19901 Germantown Road
Germantown MD 20874-1290

[redacted]

Dr Jane Peterson
National Human Genome Research Institute
National Institutes of Health
31 Center Drive MSC 2033
Building 31, Room B2B-07
Bethesda, MD 20892-2033

[redacted]

Dr Matthias Platzer
IMB Institut fuer Molekulare Biotechnologie
Department of Genome Analysis
Beutenbergstraße 11
D-07745 Jena
Germany

[redacted]

Dr Juliane Ramser
Max-Plank-Institut fuer Molekualre Genetik
Ihnestr. 73
D-14195 Berlin

[redacted]

Dr Sian Renfrey
The Wellcome Trust
183 Euston Road
London NW1 2BE

[redacted]

Dr Jane Rogers
The Sanger Centre
The Wellcome Trust Genome Campus
Hinxton
Cambridgeshire CB10 1RQ

[redacted]

Dr André Rosenthal
metaGen GmBH
Ihnestrasse 63
14195 Berlin

[redacted]

Dr Yoshiyuki Sakaki
Genomic Sciences Center, RIKEN
C/o Human Genome Center
Institute of Medical Science
University of Tokyo
4-6-1 Shirokanedi
Minato-ku
Tokyo 108-8639
Japan

[redacted]

Dr Gregory Schuler
National Center for Biotechnology Information
National Institutes of Health
38 Library Drive
Building 38A, 8N805
Bethesda, MD 20892
USA

[redacted]

Dr Nobuyoshi Shimizu
Department of Molecular Biology
Keio University School of Medicine
35 Shinanomachi
Shinjuku-ku
Tokyo 160-8582

[redacted]

Dr Douglas Smith
Genome Therapeutics Corporation
100 Beaver Street
Waltham, MA 02453-8443
USA

[redacted]

Dr Hideaki Sugawara
DDBJ
National Institute of Genetics
1111 Yata
Mishima
Shizuoka 411-8540
Japan

[redacted]

Dr John Sulston
The Sanger Centre
Wellcome Trust Genome Campus
Hinxton
Cambridgeshire CB10 1SA

[redacted]

Dr Jean Thierry-Mieg
National Center for Biotechnology Information
National Institutes of Health
38 Library Drive
Building 38 N-805
Bethesda, MD 20892

[redacted]

Dr Robert Waterston
Genome Sequencing Center Box 8501
Washington University Medical School
4444 Forest Park Avenue
Box 8501
St Louis, MO 63108
USA

[redacted]

Dr George Weinstock
Baylor College of Medicine
One Baylor Plaza
MSC-226
Houston, TX 77030

[redacted]

Dr Jean Weissenbach
GENOSCOPE
Centre National de Séquencage
2 Rue Gaston Crémieux, CP 5706
91057 EVRY Cedex
France

[redacted]

Dr Richard Wilson
Genome Sequencing Center, Box 8501
Washington University Medical School
4444 Forest Park Avenue
St Louis MO 63108

[redacted]

Dr Huanming Yang
Human Genome Center (Beijing Center)
Institute of Genetics
Chinese Academy of Sciences
Beijing, 100101
China
☎ [redacted]
Fax:
[redacted]

Updated September 12, 2000

AGENDA
Eighth International Strategy Meeting on Human Genome Sequencing
GENOSCOPE: Evry, France

<u>September 14, 2000</u>

8:00 p.m. Reception in the Restaurant of the Hotel Mercure

<u>September 15, 1000</u>

8:00 a.m. Continental breakfast in the "Francois Jacob" meeting room at Genoscope

8:30 a.m. WELCOME: Jean Weissenbach on behalf of the French Genome Project

SESSION I: ANALYSIS OF THE WORKING DRAFT AND PREPARATION FOR THE PUBLICATION

8:45 a.m. ANALYSIS - Co-chairs: Eric Lander & John Sulston

- review efforts of the Hard Core Analysis group to build paths through the data and analyze its content - David Haussler, Jim Kent, Ewan Birney, and Greg Schuler
- what remains to be done to the data for the next freeze

10:00 a.m. Coffee Break

11:30 a.m. PUBLICATION PLANS - Chair: Francis Collins

- accompanying manuscripts
- Celera plans for simultaneous publication

12:00 p.m. Lunch

SESSION II: FINISHING

1:00 p.m. Chair: Rick Wilson

- plans for coordination
- timeline for finishing the Human Genome, including current projections for the completion of individual chromosomes

SESSION III: TRACE REPOSITORY

2:30 p.m. Co-chairs: Jean Thierry-Mieg and Ewan Birney

- nomenclature - Harley Gorrell
- plans and timetable for establishing the repository
- discussion of de-archiving human traces

3:00 p.m. Coffee Break

SESSION IV: FUTURE DIRECTIONS

3:15 p.m. Chair: Michael Morgan

- establishing an International Large-genome Sequencing Consortium

4:30 p.m. Summary and Conclusions

Center	# reads/kb for full shotgun	av. # of gaps/100 kb remaining in full shotgun	av # of reads in current draft coverage	single or double-ended for draft	# of additional reads/kb to bring draft to full shotgun	single or double-ended reads for rest of full shotgun	# Mb of working draft that can be topped by by July 2001	# drafted clones that can be brought to full shotgun by July 2001
BCM	21	2	10,5	95% single 5% double	10	double	243	1700
Beijing	25	2	15	double	10	double	30	200
Genoscope	*	*	*	double	*	double	80**	540 -550**
GBF								
GTC	18,5	8	11,5	double	7	double	45	266
IMB	20 - 25	4	13	single	7 - 12	both (double for new libraries)	6	48
ISB	16 - 18	2 - 3	11	double	6 -8	double	22	150
JGI								
Keio	14.5 (8 coverage)	7	14,5	double	no need	double	no more needed	none
MPIMG	20	5	8 - 9	double	12	double	3,5	30
RIKEN	20	3 - 5	7 -10	double	8 - 10	double	100	700-800***
Sanger Centre	20	3 - 5	8 - 9	double	15	double	~600	4000
Stanford	23	2,3	22	single	3,4	single	11,6	77
Wash U	22 - 26	3,5	12 - 14	single	12 - 14	double	400	3000
WIBR	25	1,9	11	single	14	double	900	5400
UWGC	25	08-juin	~4X	both	12	double	100	840

* 520 clones are already at full shotgun coverage, and finishing has been undertaken for practically all of these. About 80 BACs are left that are not fully gap-filled; many of these are in 2-3 contigs, the rest are in up to 10 fragments.

** by December 31, 2000

***we would like to take a nested deletion strategy at the finishing stage

Post-working draft efforts: Finishing the human sequence

September 15, 2000

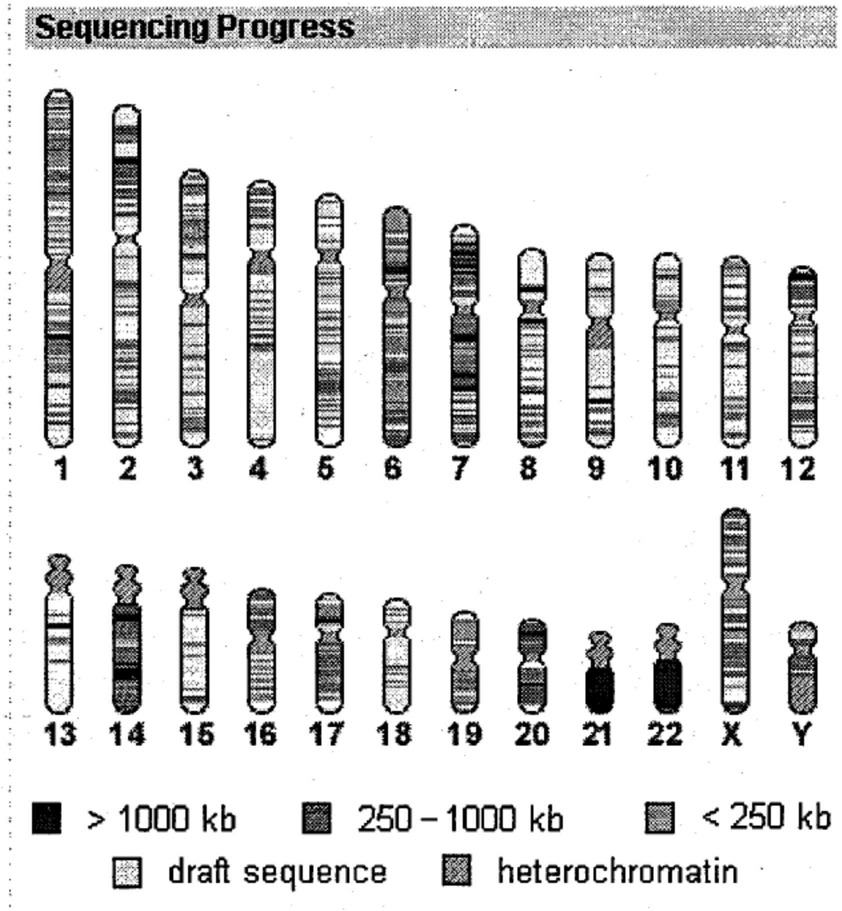
Topics

- What needs to be accomplished?
  - Top-off & finishing
  - Direction & management
- Specific concerns
  - G16 top-off capacity & finishing plans
  - Chromosome claims: "Turf wars"?
  - Continuing/future role of the Finishing Working Group
  - Ongoing monitoring of progress & sequence quality
  - Mechanism for turf redistribution
  - Standards & vocabulary
  - Resource recommendations

What needs to be accomplished?

- "Top-off" of half-shotgun projects
  - Approximately 5x additional sequence coverage (plasmid subclones)
  - Center-to-Center transfer of primary data from draft coverage
  - Trace repository
- Finishing to uniform standards
  - Quality
  - Contiguity
  - Annotation

What needs to be accomplished?



What needs to be accomplished?

- Direction & management
  - Initial assignment & coordination of “turf” claims
  - Mechanism for monitoring of progress
  - Mechanism for redistribution of “turf” when necessary
- Adherence to standards
  - Finishing Working Group
  - QA exercises

- Top-off target date = June 30, 2001
- Finished target date = April 25, 2003

Role of the G16?

- **The bulk of the finishing ☆**

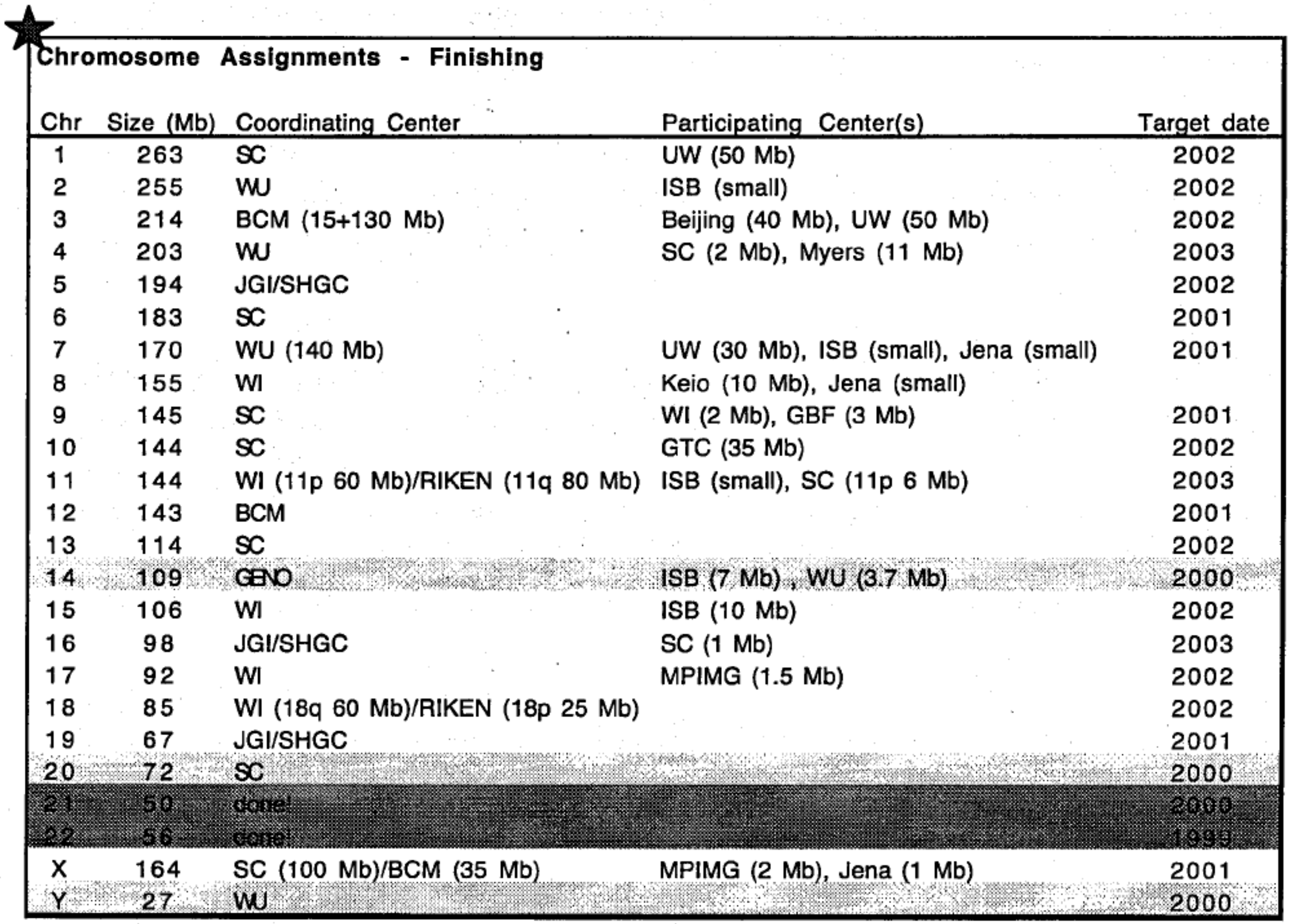
PROJECTED G5 COMBINED FINISHING EFFORT 2000-2003

Mb/Month	Sanger	WashU	WI	Baylor	JGI/Myers	TOTALS
Jul-00	20	9	4	2.5	6	41.5
Aug	23	10	4	3	6	46
Sep	25	11	4	3.5	7	50.5
Oct	25	12	4	4	8	53
Nov	25	14	4	5	8	56
Dec	25	16	4	6	8	59
Jan-01	25	18	4	7	9	63
Feb	26	19	4	8	9	66
Mar	28	20	10	10	9	77
Apr	30	21	10	11	9	81
May	30	22	20	12	9	93
June	30	23	20	13	10	96
July	35	24	30	14	10	113
Aug	35	25	30	14	10	114
Sep	35	26	40	14	10	125
Oct	40	27	40	14	10	131
Nov	40	28	40	14	10	132
Dec	40	29	40	14	10	133
Jan-02	40	30	40	14	10	134
Feb	40	31	40	14	10	135
Mar	40	32	40	14	10	136
Apr	40	33	40	14	10	137
May	40	34	40	14	10	138
June	40	35	40	14	10	139
July	40	36	40	14	10	140
Aug	40	37	40	14	10	141
Sep	40	38	40	14	10	142
Oct	40	39	40	14	10	143
Nov	40	40	40	14	10	144
Dec	40	40	40	14	10	144
Jan-03	30	35	40	14	0	119
Feb	30	35	40	14	0	119
Mar	30	35	40	14	0	119
Apr	30	35	40	14	0	119
Totals	1137	919	952	393	278	3201.8

Role of the G16?

- The bulk of the finishing ☆
- Direction & management ★
  - Coordination of "turf" claims
  - Monitoring of progress
  - Redistribution of "turf" when necessary

“Turf wars”?



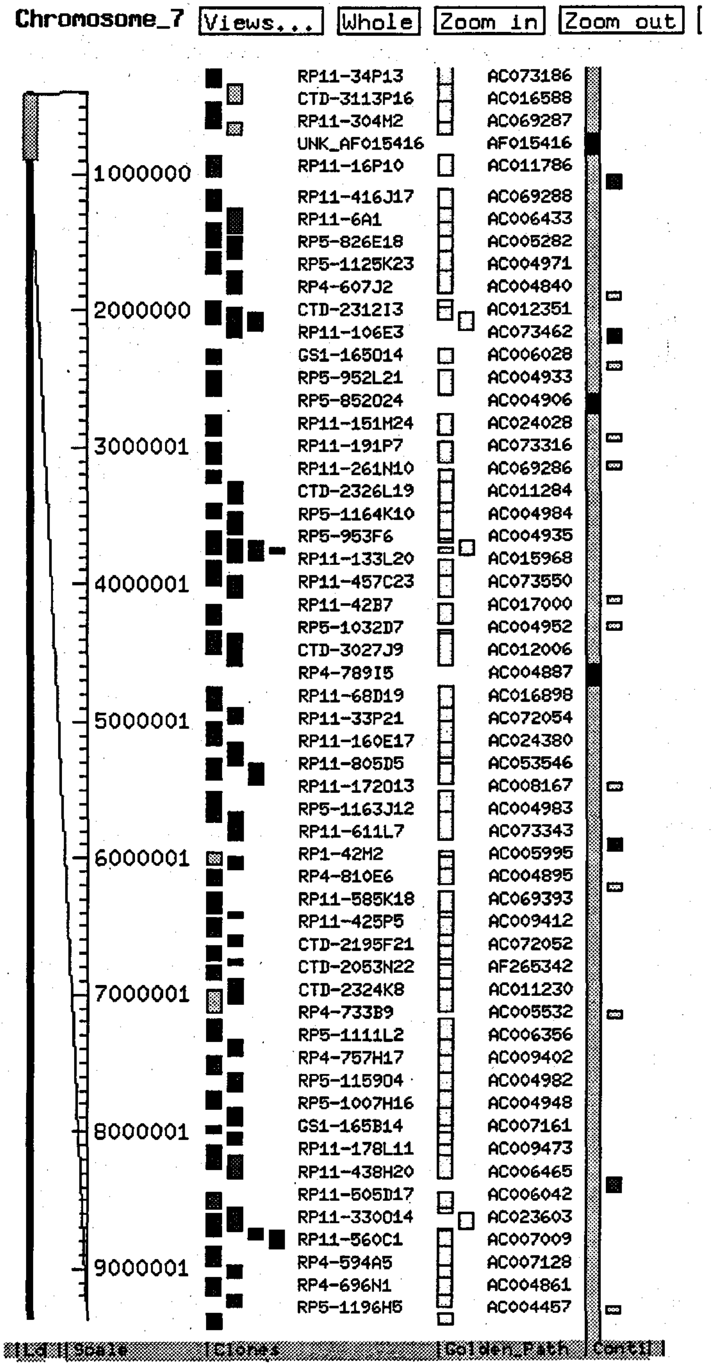
★Chromosome Assignments - Finishing

Chr	Size (Mb)	Coordinating Center	Participating Center(s)	Target date
1	263	SC	UW (50 Mb)	2002
2	255	WU	ISB (small)	2002
3	214	BCM (15+130 Mb)	Beijing (40 Mb), UW (50 Mb)	2002
4	203	WU	SC (2 Mb), Myers (11 Mb)	2003
5	194	JGI/SHGC		2002
6	183	SC		2001
7	170	WU (140 Mb)	UW (30 Mb), ISB (small), Jena (small)	2001
8	155	WI	Keio (10 Mb), Jena (small)	
9	145	SC	WI (2 Mb), GBF (3 Mb)	2001
10	144	SC	GTC (35 Mb)	2002
11	144	WI (11p 60 Mb)/RIKEN (11q 80 Mb)	ISB (small), SC (11p 6 Mb)	2003
12	143	BCM		2001
13	114	SC		2002
14	109	GENO	ISB (7 Mb) , WU (3.7 Mb)	2000
15	106	WI	ISB (10 Mb)	2002
16	98	JGI/SHGC	SC (1 Mb)	2003
17	92	WI	MPIMG (1.5 Mb)	2002
18	85	WI (18q 60 Mb)/RIKEN (18p 25 Mb)		2002
19	67	JGI/SHGC		2001
20	72	SC		2000
21	50	done!		2000
22	56	done!		1999
X	164	SC (100 Mb)/BCM (35 Mb)	MPIMG (2 Mb), Jena (1 Mb)	2001
Y	27	WU		2000

“Turf wars”?

Chromosome Assignments - Sep 00

SC	900
WU	605
WI	465
SHGC	350
BCM	323
UWGC	130
Genoscope	109
RIKEN	105
GTC	35
Beijing	30
TOTAL	3052



RP11-34P13	AC073186
CTD-3113P16	AC016588
RP11-304M2	AC069287
UNK_AF015416	AF015416
RP11-16P10	AC011786
RP11-416J17	AC069288
RP11-6A1	AC006433
RP5-826E18	AC005282
RP5-1125K23	AC004971
RP4-607J2	AC004840
CTD-2312I3	AC012351
RP11-106E3	AC073462
GS1-165O14	AC006028
RP5-952L21	AC004933
RP5-852O24	AC004906
RP11-151M24	AC024028
RP11-191P7	AC073316
RP11-261N10	AC069286
CTD-2326L19	AC011284
RP5-1164K10	AC004984
RP5-953F6	AC004935
RP11-133L20	AC015968
RP11-457C23	AC073550
RP11-42B7	AC017000
RP5-1032D7	AC004952
CTD-3027J9	AC012006
RP4-789I5	AC004887
RP11-68D19	AC016898
RP11-33P21	AC072054
RP11-160E17	AC024380
RP11-805D5	AC053546
RP11-172O13	AC008167
RP5-1163J12	AC004983
RP11-611L7	AC073343
RP1-42M2	AC005995
RP4-810E6	AC004895
RP11-585K18	AC069393
RP11-425P5	AC009412
CTD-2195F21	AC072052
CTD-2053N22	AF265342
CTD-2324K8	AC011230
RP4-733B9	AC005532
RP5-1111L2	AC006356
RP4-757H17	AC009402
RP5-1159O4	AC004982
RP5-1007H16	AC004948
GS1-165B14	AC007161
RP11-178L11	AC009473
RP11-438H20	AC006465
RP11-505D17	AC006042
RP11-330O14	AC023603
RP11-560C1	AC007009
RP4-594A5	AC007128
RP4-696N1	AC004861
RP5-1196H5	AC004457

Working Group

- Standards & vocabulary

  genome.wustl.edu/genome/fin rules

Working Group

- Resource recommendations
  - Paired end plasmid clone reads
  - Data exchange
  - Trace repository
  - Database requirements

Role of the G16?

- G16 meeting - 9/14-15/00
  - Finalize initial “turf” claims
  - Ensure top-off capacity & target date
  - Approve plans for progress & quality monitoring

Role of the G16?

- The bulk of the finishing ☆
- Direction & management ★
  - Coordination of "turf" claims
  - Monitoring of progress
  - Redistribution of "turf" when necessary
- Adherence to standards
- Finishing Working Group
  - ✓ Standards & vocabulary
  - ✓ Resource recommendations
  - Coordination of "turf" claims
  - Monitoring of progress & quality

Working Group

Donna Muzny - Baylor

Rick Myers - JGI

Stephan Beck - Sanger Centre

Chad Nussbaum - Whitehead Inst.

Rick Wilson - Wash U. (chair)

John Sulston - Sanger Centre (ad hoc)

Jane Lamerdin - LLNL (ad hoc)

Adam Felsenfeld - NHGRI

Working Group

- Collaboration & communication
  - Coordinate finishing on a per chromosome basis
  - Early detection of impedance mismatches
  - Redistribution of "turf" and effort

CONFIDENTIAL

September 9, 2000

Genome Paper Outline

(Still Rough!)

1. Introduction

Broad Historical context: Sweep of 20th C – starting with revival of mendal in 1900 to DNA to Genome Project

Here we report the results of an intl collaboration to generate 'draft' sequence covering the vast majority of the human genome. Map has clones covering 99% of the genome, have generated sequenced from clones covering 93% of the genome, The sequence itself covers 87% of the genome. How much finished, near-finished, draft, genome-WGS.

The seq is very large – 1000x larger than a typical bacteria, 25X larger than any other organism sequenced; 10X larger than sum of all previous genomes.

The sequence is not complete, but the task of finishing at least the portion of the genome in the clones is straightfoward and expect finished or near-finished state for >95% during the next year.

However, even without the gaps filled in, the data represents a huge trove of information valuable. Our goal in this paper is to take a first look at the genome, give a basic description, lay out main results, issues and agenda. Although the precise numbers will surely change with additional data and analysis, the main results already emerge:

They paint a picture of: repeats, genes, isochore neighborhoods, particular gene families.
It is a history book, identify the gene families (modest number of genes)

2. History of Genome Project

The proposal to sequence the human genome was first floated in discussions at ... in 1984-6, touching off controversy. From those discussions emerged a richer project, with a number of intermediate goals-genetic maps, physical maps, other organisms, ethics, technology, etc.

Efforts were launched in a number of countries, -- by govt funding agencies and foundations and universities and non-profit research insitutions. Led to the production of many many things

The actual task of sequencing the human was divided into a two phases: pilot phase 1996-1999 and production phase from 1999 onward.

2.1. Pilot Phase

Goal of the pilot phase to address:

2.1.1.Feasibility – address issues of accuracy, coverage – would it be possible to produce long-range continuous sequence, what types of artifacts. Best way to address this was to study individual large insert clones (cosmids and BACs to be able to focus on human DNA). Developed way to check that the assembly is correct by comparing to FP.

Studied how coverage is obtained with half-SG; full SG; finishing—worked out this method.
Using M13 vs plasmid
Captured gaps etc.
In extreme case, also looked at complete chromosomes

2.1.2. Efficiency – how to to this efficiently. Reduce costs from $2-5/finished base or about raw read of $20?

- Development of Automation (PICTURE of WIBR machines), similarly drove sequence detectors
- Computer Programs (Phred, Phrap, etc)
- Data Production – Produce a modest amount of the actual sequence (how much was called for), driving analysis

2.1.3. Collaboration – Although it was recognized that the maximal efficiency would likely be achieved by a small number of large factories (wrt economies of scale, ease of coordination, QC), it was felt that there was an inherent value in broad involvement.

Forge an intl consortium – with different countries

Some groups took respons for individual chromosomes, others more whole genome. Coordinate map.

Used a diversity of expt methods

2.1.4. Data sharing –Pioneering models of pre-production data release

2.2. Production Phase: Choice of Strategy

2.2.1. Hierarchical SG vs WGS. The key issue for large-scale production was:

All sequencing is performed by SG sequencing – but the issue is how large should the segment be.

Continue with clone-by-clone approach or adopt a WGS shotgun approach.

Potential advantages of each (save on the overlap, can get regions that cannot clone in BACs)

SG sequencing developed by Sanger—trivial from math standpoint to assemble as long as unique.(Rapid assembly)

There has been a steady increase in size of segment

Decided that the nested SG made the most sense:

- Existence of nearly perfect repeat segments
- Desire to end up with a eventual product that was as close to perfect as possible
- Lack of a method to spot the misassemblies in a WGS
- Feasibility of generating FP coverage
- best for coordinating an intl project. Let's everyone take independent responsibility and pride too.
- allows haplotype reconstruction, avoids issues of structural heterogeneity (cite such places)

For all these reasons, decided that do hierarchical SG

Happily, Celera Genomics decided to take a different approach provides a complementary. Will give an examine the issues in the context of data

The correct strategy for other organisms (may depend on degree of repeat identity on genome structure – size of repeat blocks, activity of transposable elements (which controls % identity)

2.2.2: Emphasize Draft first.

Once we had demonstrated that one could reliably finish BACs, made sense to cover as much as possible first.

Decided to break at Draft Phase, with some finishing.

To be followed by Near-Finished Phase,

To be followed by Finished Phase

3. Data Production: Seq, Assembly, Evaluation

3.1. Clones selection and sequencing

3.3.1. Sequencing of clones.

Clones selected in a variety of ways (see other papers of map)
Subjected them to SG – either half-, deep- or finished,
M13 vs plasmid

Target date was end of June, most of the data we report here was produced during that period, continues to grow

Descriptions:
How many are draft? Etc.
How many-fold cov of the genome?

Are draft contigs O+O? no. How many contigs, how much of the clone is in contigs of what size? As we discuss below, the accuracy is high and the gaps are .....

3.3.2. WGS Sequencing

•• Also generated WGS reads for about 50% coverage
Chosen for SNPs too

3.2. Assembling the Genome Sequence

3.2.1. Cleaning up the sequence

Clean up clones for contamination, mislabelling, merged projects
Necessary because a variety of centers using different techniques

3.2.2. Creating the GP:

Merge them into a best-guess path
Accessions initially assigned to FP contigs
Also assigned FP contigs to genome by genetic/RH map
Thus the physical map provided an essential scaffold as planned

Check that clone did overlap by Seq with clones that it ought to overlap
Check that STSs contained in the clone (Describe RH content)
On basis of this, could identify clones that did not belong – any interesting cases to describe??
Clones assembed into C-contigs, SC-contigs, S-contigs
Merging may not be perfect
Use of the O+O reads

3.2.3. Description of the GP

Describe the GP
Single integrated view of the genome
Discuss bridging

3.3. Quality Assessment

3.3.1. Quality assessment—indiv clones

Of indiv clones of different degrees of covg
Missassemblies of indiv clones

3.3.2. Quality assessment—of GP

Phrap scores
Recheck for contamination
For erroneous STSs
What did we do with misassemblies

3.4. Coverage of the Genome

Coverage assessment

What fraction of individual clones is missing?

Can we look at how much is missing by looking at the graph of gaps vs coverage and then looking at coverage in the GP
What is the number of gaps in finished sequence

4. Genomic Landscape

Protein-coding genes comprise only a tiny fraction of the genome.
We focus instead on large-scale features of the composition of the genome.
Repeats-thought of as junk, but contain extraordinary archeology

4.1. GC content

Plot of the CG content of windows of size X (How does this vary with X)
Plot of CG content of individual chromosomes?
How big is an isochore as a function of its CG content?

Can we write an HMM to find CpG islands?
What are some extreme cases of CG content?
What does the mosaic look like if we mark the isochores in different colors?

Mention Extreme cases

4.2. Transposable Elements

4.2.1. Review of four classes

SINES, LINES, LTR, DNA transposons
Discuss the subfamilies.
Scanned the genome for known repeat families.

4.2.1.1. Identification of new transposons

As genome sequence has been piling up, we have been regularly updating the families. Still, we reviewed the genome to look for new families. Results

4.2.2. History of transposition

For each type of transposible element it is possible to assemble them into clades of related elements and date them, with seq similarity being a proxy for date of transposition. We are in process of reconstructing trees now, but ….[Age distribution – compare the age distribution for mouse]

•• Age distribution and clade recalculation
Remarkably quiescent

Compare to other species. – if the frequency of a clade is 1%, we ought to be able to recognize it in a modest amount of sequence. Good way to build evolutionary trees with modest amount of WGS sequence?

4.2.3. Distribution of Repeats

4.2.3.1. Distribution of elements by isochore

(minus the element itself—ie masked sequence)
What explains this distribution?
Is there a similar distribution for processed pseudogenes?

4.2.3.2. Distribution of young Alus with respect to isochore

4.2.3.3. Distribution by chroms corrected for isochore

What does it mean?
Chr 19, Chr Y. What funny repeat distributions?

4.3.3.4. Extremely repeat-dense and repeat-poor regions.

Distribution of Hox is amazing [Figure]

Hox is an extreme case in that there is selection against all repeats, but there is general selection!!!!!!

4.3.3.5. Distribution within genes

No strong selection against Alus in genes
Density of Alu in + strand of intron
Density of Alu in – strand of intron – both referenced to genomic average
For different isochores – This also can show preference of Alu for introns!!!
Dots plotted---

4.2.4. Mutational events

4.2.4.1. Transposons as a cause of mutation.

What is the freq of new hops?
Frequency of transposition and transduction of 3'-ends

4.2.4.2 Transposons as a marker of mutation

Transposible elements as a measure of local mutation rates

4.2.5. The Young and The Restless.

4.2.5.1 Identification of polymorphic elements

4.2.5.2. Identification of currently active elements.

4.3. Simple Sequence Repeats

4.4. Chromosome Organization: Centromeres and Telomeres

4.4.1. Structure of centromeres, (divergence – 96%)+satellites etc

4.4.2. Identification of ancient centromere

4.4.3. Structure of telomeres (divergence = ??)

4.5. Repeated Genomic Segments

(within the past N million years—what is N?)
Role of these things in disease

4.5.1. Small Dispersed elements (eg Chrom specific repeats)

Good number of copies, high divergence

4.5.2 Large segments

4.6 Distribution of Recombination

5. Properties of Genes

Before we turn to the challenge of creating of complete human gene index, we studied the properties of genes, ESTs, ecores. Allows us to learn a lot about ...

5.1 Studying known 'FL'-genes.

Many human genes have been well studied by various biological techniques. The availability of genomic sequence lets us understand the structure and distribution of genes by re-evaluating the known genes.

5.1.1. Unsplicing
Dataset: Description of 'known' gene sets
Table of genes: Name, length of CDS, present in genomic? Present in finished?, number of exons, length of first coding exon, length of last coding exon, lengths of internal coding exons, avg length of internal coding exons, (If in finished:) number of intron, length of introns, total genomic length from first to last coding exon, CG content (how to define for draft, for finished), expression information (house-keeping, hi expression)
Figures: Scatterplot of number of exons vs CG; total length vs CG; average intron size vs CG;
Facts: Genes with most exons, largest exons, longest genomic length
The introns should be sent to Arian and evaluated for their repeat content.
Table: Distribution of known genes by chromosome

5.1.2. Processed pseudogenes of known genes
How many are there? Where are they distributed? Do they show the same CG bias as LINEs (which is what you might expect given that they hop using the LINE machinery)

5.2 Re-evaluating the EST collection

Table: Numbers of PCECs
Figures: Distribution by chromosome

Another approach is to use ESTs. The availability of genomic sequence lets us re-evaluate the EST collections.

What is the representation factor R? (R= fraction of known genes with a PCEC)
What is the inflation factor K? (K= number of PCECs that derive from same gene)

If N denotes the number of PCECs, then our estimate of gene number is N/RK.

Assumes that the known genes are not more or less likely to have PCECs – ie not biased to be in the EST collection.

5.3. Ecores

5.4 CpG Islands

What fraction of known genes have CpG islands? How many CpG islands are there?

6. The Human Gene Index

challenge of finding human genes etc.
include distribution of gene density, clustering of genes
largest genes in the genome etc.

6.1. Methodology for gene finding

6.1.1 Structural RNAs

6.1.2. Protein coding genes

6.2. Description of the Human Proteome

6.3. Gene Clusters

6.4. Comparative Proteomics of Eukaryotes

6.5. Stories about genes

6.6. Toward a complete proteome gene set

6.7 Beyond just the genes: Regulatory regions

7. Ancient History: Syntenic Regions between and within species

7.1 Human-Mouse Synteny

Any synteny with fly?????

7.2 Ancient Duplications in the Vertebrate Lineage: Octoploidy

Genome History
•• Ancient Duplicated segments
•• Synteny; Regulatory regions

If we can still line up LINES from before the mammalian radiation, why cant we line up genomic DNA????

8. Recent History: Human Polymorphism

SNPs

9. The Road Ahead: Next steps

10. Applications, Implications

Uses that have been made of the data so far

Uses that will be made

CONFIDENTIAL

Updated September 10, 2000

Genome Paper Analyses

This document outlines the analyses that are currently underway (as of early September) in preparation for the main sequence paper.

1.The Sequence

1.1. Individual Accessions **Adam Felsenfeld**

Large-insert libraries. For each library:

- Library name,
- Clone type (BAC, PAC, cosmid), type of source DNA, enzyme digest
- Average Insert Size of Library
- Total Number of Clones in Library
- Number of Clones Successfully Fingerprinted
- Number of Clones Successfully End-Sequenced (at at least one end)

Sequence Accessions

- Number of Clones in each Sequence Category (Fin, Draft, Pre-Draft, Skimmed). [Remember to eliminate the 3 kb plasmids; only cosmids and above]
- Number of Total Mb in each category
- Estimate of average-fold coverage, in each category (so we can report total Gb of data)

1.2 The Golden Path **Jim Kent**

(Text) Description of procedure for making the GP from the clone map

---

1.3 Genome Coverage Analysis **Eric Lander**

Described in separate memo (attached). Key points:

- Resampling estimates
- Experimental evaluation of no-hitters and gap sizes
- Key tables describing properties of the GP

1.4 Accuracy Analysis **John Bouck**

Described in separate memo. Key points:

- (Table, Graph) John Bouck to provide number, distribution of Q-scores in finished, in draft, in pre-draft

• (Table) Greg to repeat missasembly analysis using all intersections of draft with finished. Keep score (of both numerator and denominator in contigs, bp) as a function of Q-base coverage.

• (Tables) Greg to perform analysis on conflicts between GP and RH position

•••Also, what are the missed overlaps in the GP? Other errors in the layout?

1.5 Finishing State of the Genome **David Kulp**

1. (Chromoplot) Display the finished state showing finished, deep SG and draft.

1.6 Isochore Analysis **David Haussler**

(David Kulp will bring DH up to speed)
(help from James Galagan)

Goal 1: We want to perform a careful analysis of the "isochore" structure of the genome. In particular, we'd like to know in what sense we can meaningfully discuss the genome being "divided" into discrete isochores.

We have defined this as the persistence of GC ratios over substantial distances. If sequence was randomly chosen with Pr(CG)=0.42, then the CG-fraction in non-overlapping segments would be completely uncorrelated. But, there is clear correlation. We want to know how far it extends and whether we can divide the genome into "blocks", separated by change points.

Suppose that CG-content changes as follows: Sequence arrives with a GC-content Y until one reaches a change-point, after which sequence arrives with GC-content Z. Change points follow a random arrival process with characteristic distance a, and Y and Z are independently drawn from a CG-distribution D(X) (and thus uncorrelated).

Analyses:

1. **CG-distribution**. Produce a frequency distribution of the CG-content of the genome averaged over a suitable window size. (Investigate the effect of window size). Provide mean, median, distribution (in Excel spreadsheet) and a graph.
2. **Autocorrelation function**. Produce the autocorrelation function, giving the correlation C(x) of CG-content at points separated by distance x. (Investigate the effect of window size, use of finished versus draft; ways to deal with gaps in draft, including never reaching over sequence gaps. Study this on Chr 21 and Chr 22 to examine cases where we have truly long range information.)
   [Note: For random arrivals of changepoints, the autocorrelation function is simply Exp(-x/a). Accordingly, we can examine the shape of the function to see if it is log linear and estimate the constant a.
   We can also estimate the slope of the autocorrelation as a function of the CG content. This may reveal to use that different CG content blocks have different persistence lengths
3. **Find the change-points**. Assuming the process above, we can write a simple HMM that estimates the "hidden" CG-content at each position of the genome as well as the location of the change-points. We need to give the HMM the value of a and the distribution from which to draw the CG-contents. We can use the estimate of a above (as well as trying others; it should not be too sensitive provided that we give it rather discrete choices for CG-content; we might try letting it have 4-5 values to use). We should try this on Chr 21 and 22 to see what happens. If this is successful, we might want to mark up the genome in this fashion.
   Proposal for HMM model: Five states (S1-S5) and five characteristic transition lengths (T1-T5). Estimate the Markov transition probabilities and the lengths. How much better does this do than single T? How many transition points? What are the properties?

Goal 2. CG-content of chromosomes.

1. (Table, Graph) For each chromosome, proportion of sequence in each isochore. [Arian has this: Shows 19, 22, 17, 16 as clear outliers]
2. (Chromoplot—Neomorphic) Distribution of CG-content across chromosomes. [David Kulp will explore altering the chromoplot viewer to "spikes" as in C. elegans paper.]
3. [Table] Get the table of relative gene density as reported in the 30,000 gene RH paper.
4. Describe example of the most CG-rich and CG-poor regions, over some substantial distance.

Goal 3: We are thinking that the differences between the "two genomes" may be an important theme. We want to understand how different properties vary with the "isochores" or at least with the GC-content. These properties include:

- Density of different repeat families
- Density of genes
- Gene structure (number, sizes of exons/introns, initiator codons, splice sites, ...)
- Gene type (housekeeping vs tissue specific, preference of specific genes families)
- Recombination rates
- Nucleotide substitution frequencies
- Chromosome banding patterns

We will discuss these below.

1.7 Recombinational Distance vs Physical Distance David Kulp

We think that the gapping problems in the Golden Path probably smooth out at the level of 1-2 Mb. Accordingly, we may be able to compare recombinational and physical distance despite the gaps.

1. Map the Genethon markers onto the GP. For each chromosome, plot the physical distance along the chromosome vs the physical distance.

David Kulp will make the tables of chrom x chrom.
ESL will contact Jim Weber and see what's going on?

1.8 rDNA: Relevant to Genome Coverage Victor Pollara

1. Where are the ribosomal RNA genes? We don't have many in our accession. [These are the acrocentric short arms. They are not heterochromatin.]
2. How many of the WIBR WGS reads contain rDNA? (Even a single copy of the repeat should count as a hit in this analysis)

CONFIDENTIAL

Updated September 10, 2000

2. Repetitive DNA

2.1 Known Transposable Elements Arian Smit

We want to provide a comprehensive discussion of the properties of repeats. Some (many) of these analyses have been done before, but we want to repeat them on this much more complete dataset.

I. Transposon-based repeats

Basic properties of the repeats.
For each major type of repeat (SINES, LINES, LTR-based, DNA transposons), we want to know the following

1. (Text) Biology of the repeat.
2. (Table with text). Important subfamilies (E.g., The SINEs include the Alus and the ancient MIRs; the Alus have lots of different sub-families (give names of major ones and brief description. There are 250 LTR-based transposons with typical families having *** members, most are represented by solitary LTRs).
3. For a given repeat: What is the rate of fixation of new repeats? What is the actual mutation rate of new hops per generation?
4. (Table, Graph). Age distribution of repeat families. (I think it would be good to superimpose different frequency plots, showing the different peaks clearly. If different DNA transposons have tight but different age distributions, this might be shown with some example pictures that make this clear). Textual comments about how % sequence divergence correlates with age, when repeats died out, about clear difference with the mouse.
5. (Table) Density of elements: # per Mb and % of DNA overall; for individual chromosomes; and for different CG-fractions.
6. (Table) Density of repeats elements per chromosome, normalized by expected density as a function of CG-content. [Note: We want to normalize by looking at the proportion of chromosome in each isochore and the expected number of repeats in the isochore.]
7. (Table, Graph) Proportion of repeats found in +/- direction in introns in known genes in finished sequence.
8. (Chromoplot) Distribution of major repeat families across the chromosomes (We need to specify which families).
9. (Chromoplot) Distribution of major repeat families across the chromosomes corrected for expected density as a function of GC content. (We need to specify which families).

Particular distributional issues

1. (Table, Graph) Frequency of different Alu families in different isochores (normalized by overall expected Alu frequeny in the isochore).
2. (Table, Graph) Distribution of LINE lengths in different isochores (overall and by subfamilies of different ages)
3. (Table, Graph) Distribution of LINE, Alu age with isochore
4. Unusual properties of elements on sex chromosomes and what it might mean

Identification of Extremely Repeat-Rich and Repeat-Poor Regions

1. (Pretty Pictures) Some regions that are wall-to-wall repeats. (eg 500 kb of LINE and Alu on X; Satellite regions from WIBR)
2. (Pretty Pcitures) Repeat-poor regions (Ken Dewar's paper on HOX)

Currently and Recently Active Elements
It would be good to scan the sequence to identify the possible active source genes that can account for de novo mutations

1. (Text) Track down the de novo SINE and LINE insertions. How many? How similar in sequence? Which ones have already been tracked down and how? Scan the sequence to find the reasonable candidates. How many are there that could possibly have been the source?
2. (Table, Graph) Distribution of young elements by CG-content and by chromosome.
3. (Text) Arian will write up what is known about the level of polymorphism of the young classes of SINE and LINE?

Recluster the SINEs

1. (Text) We are not going to re-tree the repeats. But, Arian will write a nice description of how this can be done using the 1 million Alus and what we will learn from it.

Identification of new genes created from repeats

1. Can we find any new examples.

Mutation Rates in Different Isochores

1. (Text, Table) Mutation Rates as estimated from DNA transposon families.
   [Why does it seem that long term mutation rate is pushing us to lower CG? How do these rates compare with those estimated from new mutation data? [SNPs, or recent paper on Hemophilia B (Green, Gianelli. AmJHG 65:1572-1587 (1999). Evidence of selection on silent sites composition in mammals. Eyre-WalkerGenetics 152;675-683(1999)?]
2. Arian will re-examine the high CG isochores to see if the conclusion still holds up.

2.2 Simple sequence repeats (SSRs) John Bouck

1. (Text) Precise definition used for recognizing SSRs – both mini- and micro-satellites. For microsatellites, we should break it out by specific repeat type (eg CA or GA, at least for the major ones). Want to focus on longest track of purity being at least 13 repeats; they become polymorphic around them.
2. (Table) Density of elements: # per Mb and % of DNA overall; for individual chromosomes; and for different CG-fractions.
3. (Result). We specifically want to know why there is a deficit of polymorphic CAs on Chr X. Is it due to a deficit of CAs of sufficiently length? Or is it due to a lack of polymorphism (a population genetic consequence of being on a chromosome with a smaller effective population size).
4. Arian needs to look back at this. He found that CAs are over-represented in GC-rich, but perfect CAs are under-represented!! Hmm? Are the CAs generated by repeats?
5. Ewan will send an email to Compugen asking to see if we can get repeater.

2.3 Identification of new repeats Evan Eichler

1. (Evan/Greg) Unbiased look at genome. Take all accessions (or should we do GP?). Repeat-mask them. Gene mask them. Tandem mask them. Look at the rest. Blast against each other for 85% identity at 100 bp more than 25 repeats in the unique genome. Describe families that arise.
   [Can we be somewhat comprehensive by looking for all examples of families with at least N elements having average conservation X over length Y (for some reasonable choice of N,X,Y).]
2. (Ewan→Serafim). Using supermasked path. Identification of common words by hashing that cannot be explained by known repeats?

3. (Ewan/Arian) Using protein analysis to identify new LTRs and DNA transposons (looking for protein motifs, using WISE to pick up decayed protein motifs). Ewan will attack this with help from Arian on parameters. Arian will take a look at what Ewan finds.
4. Are there any Ty-Copias in genome? Why do we want to know? (This should be caught in section 2 above?)

(Text). We need to write up the methods used?

2.4 Satellites/Heterochromatin — Evan Eichler

1. (Text) Definition and Description of known satellite/heterochromatin repeats. What's known about how much of it is in the genome and where it is, by virtue of cytogenetics?
2. (Table). Frequency of known satellites in large-insert clones, in WGS plasmid library. How many BACs do we have with satellites? Where do they map?
3. We need to check the correct location of the clones that have these satellites.
4. Compugen has nice code, REPEATER, for recognizing repeats in a large piece of DNA.
5. (Text) Will try to look for new satellites, in known accessions with satellites.
6. (EXPERIMENT). Prepare a library of very short (100bp or so) sheared fragments and sequence 2,000-10,000 reads to look for the frequency of satellites. We can use this to estimate whether we are significantly underestimating this proportion of the genome. [**Bruce Birren** will coordinate this at WIBR]
   We should be careful to trim all reads being compared so that they are the same lengths.
7. Need to write up methods!!!!!

2.5 Large-Scale Duplications — Evan Eichler

General Information about Large-Scale Duplications

1. (Text) Considerable text describing known examples of large duplications in the genome, their role in disease, the overall level of large duplications (number of events, proportion of genome, typical size).
2. (Text, Table) Comparison of these to other organisms (yeast, fly, worm)
3. (Text) What is known about heteromorphisms in such regions?

Genome Structure near Centromeres and Telomeres

1. (Text) Describe our understanding of what is going on near centromeres and telomeres, in terms of typical size, structure, age of events. Size of blocks, spacers between blocks. Hoovering up the genome! [Percent similarity (could be conversion not age). Lots of evidence of presence/absence in telomeres.]
2. (Text and Nice Illustrations). Detailed analysis of Chr 21, 22 cen and tel in light of comparison to rest of genome to identify the component elements (A,B,C,D).
3. What other chromosomes can we analyze in this manner? Chr 7, Chr 20
   Ewan will get information on Chr 20 to Evan.

Chromosome-specific repeats

1. (Text) Describe what is known about chromosome specific repeats? How do they evolve? When did they evolve? We see them in the monkeys, even though the pieces are 99% identical – implies conversion. Keeping things in order
2. What are the implications for disease?

Identification of Recently Duplicated Segments

1. (Text, Tables) Identify duplicated segments by looking for segments of length X and are <98% identical. (Does this prevent being confused by low quality sequence?) Distinguish intra and inter chromosomal repeats.
   Evan proposes to definitely do NT x NT, NT x GP. [Evan needs NT set cleaned up for overlap.]

2. Then may be able to do GP x GP.
   [Victor/Serafim has done this; Evan needs to look at it and redirect it.
   This may point to serious problems with the GP. What proportion of missed overlaps? This needs to go into the paper in terms of accuracy.]
3. (Pictures) Identify which segments are multi-copy chromosome-specific repeats: Give examples of mapping them back onto the chromosome. This may be crucial to believe any whole genome analysis.
4. Identify duplicated segments by using WIBR random WGS reads?

Vestigial Centromeres

1. (Text) What is known about vestigial centromeres? What do we see happening at these locations in the GP sequence? How big is the sequence?
2. (Text) Human has 2p-2q fusion. 2q21 = vestigial; it is shrinking rapidly. There is a pericentromeric domain near it. There are genes crawling out of the ashes, says Evan.
3. (Text) Are there other examples – possibly on 10, 15 remants. There is a nice paper about recent fusions/fissions in the last 15 million years. Do we see any evidence?

3. Genes

3.1 Known Genes David Kulp

Goal: Characterize the structure of genes, using well-known genes as our test set.

1. (Text, Table) Define the sets of "full-length" genes available for study.
   (i) REFSEQ, (ii) mRNA with complete CDS, (iii) reliable UTRs.
   [Mostly we will be looking at coding exons]
2. (Tables, Graphs) Unsplicing genes against finished sequence. What fraction of genes (in each set used) can be found in the finished sequence? For each gene: What are: the number/lengths of the exons (noting the first and last coding exons separately) and introns? total genomic locus size? [We want tables, graphs showing the distributions] Average CG-content? What is the correlation between these properties and CG-content?
3. (Tables, Graphs) Unsplicing genes against all sequence. What fraction of genes (in each set used) can be found? For each gene: What are: the number/lengths of the exons (noting the first and last coding exons separately)? [We want tables, graphs showing the distributions] What is average CG-content? What is the correlation between these properties and CG-content?
4. For CG-content, let's look at a window of 10,000 and 50,000. Is there any difference? To deal with edge effects, take largest window provided it's not too small. Write up the method.
5. If we classify genes or gene families (in either of the previous) as housekeeping vs tissue specific (or hi vs lo expression), is there any correlation with CG content? Perhaps the best approach is to look at the average CG of InterPro.
   David will have a list of genes and CG content of neighborhood, Alex will InterPro them and see if we see correlations.
6. For long introns, can their large size be "explained" by their repeat content? (Is the size difference substantially decreased if we focus only on non-repeat sequence in the introns)
7. For long introns, is there a clear deficit of repeats in the immediate vicinity of exons – as for the metabotropic glutamate receptor?
8. Do we have a good collection of "full length" UTRs? Characterize the number of introns in 5' vs 3' UTRs.
9. (Table) Can we identify processed pseudogenes, by comparing the known (spliced) mRNAs to the genome and looking for unspliced versions? [David Kulp will look at this and note cases of unspliced versions of the gene]

3.2 Revised UniGene Set Greg Schuler

Goal: Clean up the UniGene 3'-ends using genomic sequence and map them onto the genome.

1. (Text) Description of the process for identifying putatively correct 3' ends. [Involves identifying 3'-ends that do not have templated poly A and do have a polyadenylation signal. Discuss the sensitivity of the definition for known genes? For 3'-ESTs?] We want to recalculate these on the GP, rather than on the individual accessions.
2. (Text, Table) What are the numbers of pc-ESTs and PCECs found? What is the number found if we insist on the cluster being represented by at least TWO ESTs? I think we may want to use the latter definition.
3. For known genes, what is the typical number of PCECs correponding to the gene?
4. (Table, Figures) What is the distribution of local CG content for these PCECs? How does it compare to the expected CG content for the genome?
5. (Table, Figures) How do the PCECs distribute by chromosome?
6. (Chromoplot) Density of PCECs

7. How does this set compare with Phil Green's set? We want to get Phil's set. (Phil required at least two occurrences)
8. What fraction of known genes are hit by at least one (two) ESTs? What is the average number of PCECs per known gene?
9. What fraction of genes on Chr 21, 22 are hit by at least one (two) ESTs? What is the average number of PCECs per gene?

Goal: Briefly discuss how genome is being used to clean up UniGene (this will likely be the subject of an accompanying paper, but we still want to mention). Effect on splitting and lumping of UniGene clusters

3.3 Revisiting Estimates of Human Gene Number Eric Lander

1. Check with Weissenbach concerning (i) how the number of ecores has grown with 2.7 Gb of GP (rather than the 1.2 Gb when they did there analysis) and (ii) how well the ecores did at hitting on chr 21 (which came out after their paper)
2. Check with Phil Green: What does his analysis say when repeated for Chr 21?
3. Check with John Quackenbush concerning correction to their paper.
4. Check with Sanger Chr 22 team: We should mention their experience in checking the chr 22 gene set following publication. How many new genes?

3.4 Classification system for Gene Predictions Richard Durbin

Richard has agreed to propose the classification system to be used for our gene predictions (synthesizing the best of the approaches used for Chr 22 and 21)

3.5 Defining the Human Gene Set Ewan Birney

Goal: Define the Human Gene Set, as well as possible—including a discussion of how well we are doing. The tests of how well we are doing are:

(i) How often do we hit at least PART of the gene?
(ii) What proportion of the coding bases do we hit (and what proportion of the predicted bases don't hit coding bases)?
(iii) How often do we get the gene perfectly – or very close to right?

Do our predictions do better in finished vs draft sequence? (That is, is the problem gene prediction per se or is it that we have draft sequence? If we don't do significantly worse in draft, that's important to know.)

Three Methods. We have three different versions of this.

(i) ENSEMBL peptide set
(ii) NCBI gene set (using GenomeScan)
(iii) Neomorphic Gene Set

1. (Text, Tables) For each method, we want a comprehensive description of the methodology and the results—"including the different categories of gene prediction". This includes how many predictions, how large the gene predictions are (in number of exons and in bp). We should distinguish the results for finished vs draft.
2. (Table, Text) For each method, how well does it do on the known gene set? (For this purpose, we don't want to use these RNAs. Perhaps this analysis will just be restricted to a known region)
3. (Table, Text) For each method, how well does it do on Chr 21/22 gene sets? (For this purpose, we want to exclude RNAs found by special attention to this chromosome.?)

4. (Table, Text) How do the methods compare with each other? (Ewan will compute a table) Need to describe how this will be done (need to compare against locations in the genome).
5. (Tables, Graphs, Chromoplots) How do the gene predictions distribute across different CG-content, across different chromosomes?
6. (Tables, Graphs, Chromoplots) How does gene density (corrected for CG content) vary across chromosomes? Use the PCECs and the three methods.

Gene Density

1. Describe large regions of that are extremely gene-rich or gene poor, to give a sense of the distribution across the genome? For this purposes, consider known, EST, prediction.

Related Question:

1. (Table, Text) What fraction of known genes are hit by Tetraodon genomic? What fraction of genes on Chr 21/22 are hit by Tetraodon? If we had a lot of Tetraodon, what proportion of the genes/exons would we find? (Richard Durbin will sort this out.)

Ongoing Curation of Gene Set

1. (Text) How will the International Gene Index be produced and maintained? What will it consist of? Using the gene index to prioritize finishing?
   THIS IS AN IMPORTANT, WORTH MULTIPLE PARAGRAPHS!!!!!

Special Cases of genes that can be identified with high reliability

1. (Text) Identification of paralogues for the full length mRNAs by comparing them to the genome? How many human genes can we identify that, while not known full length RNAs, are long matches to them? **[David Kulp can look at this]**
2. (Text, Table) How many olfactory receptors? We want to count this. We currently have only 222 ORs; there must be 1000. We need to find them.
3. Estimates of kinases say that we should have 1100 kinases—but we only have 400 kinases!!!! Where are the rest?
4. What other gene families are we missing large numbers of? We really need to run GENEWISE on this.
5. For what other gene families can we reliably identify the members of the gene families across most of their length by structural comparison?

3.6 Nature of Human Genes/Comp. Genomics Alex Bateman

Goal: Describe the families in the human gene set and how they compare to other organisms.

1. Major Tables, Figures
   - Top 100 InterPro families
   - Interesting Expansions in Vertebrate (esp relative to other species)
   - Functional Categories (pie charts)
   - Number of paralogues in human for invertebrate genes

2. Alex's BLAST tables. How many genes in X have a strong BLAST hit in Y? in Y,Z and W?

This is described in more detail in memo on Genes/Proteins.

Identifying Local Gene Clusters

1. How to do this?

3.7 Stories about Genes and Gene Families Chris Ponting

(Text) Many, many nice stories describing interesting questions about human gene families.

3.8 Ancient Duplicated Segments S. Batzoglou/J.Galagan

Goal: Identify segments representing ancient chromosomal duplications, by virtue of containing multiple paralogues in the same order (or nearly the same).

1. One approach is to identify paralogues (such as the tetralogues set, and examine the local region of a few Mb for other known or unknown paralogues. This can convince us that we have seen a duplicated segment evident from protein but not DNA analysis.
2. We may or may not want to attempt to date these. We could examine the index paralogue (the one that we started with) and estimates its divergence age, and then see if the others agree. In the case of tetralogues, we might want to see if we have a 2+2 branching order.
3. A second approach is to (more automatically) identify all paralogues and look for duplicated segments.

This analysis tracks that done by Ken Wolfe for yeast. Ewan will contact him

3.9 Human-Mouse Synteny Deanna Church

1. For each human gene, can we identify the mouse orthologue? NCBI has HOMOLOGENE project. There are 8500 human, mouse, rat.
2. (What is a good test for human mouse orthology?) Look up the position on the mouse gene map.
3. (Pretty graphic) Synteny picture: Colored picture showing the chromosomal location of mouse homologues (in diff colored blocks) onto the GP

3.10 Path Forward: Gene Closure and Beyond! Richard Durbin

Closure on the Gene Set
We need general text describing the wisdom gained by the Sanger projects.

1. What is the best way to get to identifying remaining genes?
   (i) Middles of random ESTs? What did the Brazilian group do? What was its effect on UniGene?
   (ii) What will mouse, Tetraodon do?
   (iii) Role of full length mRNAs?
2. What is the process of Gene Validation?
3. What ongoing coordinated project will drive us to Gene Closure?

Regulatory Regions

1. What have we/will we learn from cross species analysis?

4. Other

4.1 Structural RNAs — Sean Eddy

1. Run profiles through the genome to identify instances of known structural RNAs in the GP. Which ones can he find?.
2. Examine any interesting pattern in their distribution.

Eric Lander will contact Sean Eddy.

4.2 CpG Islands — John Bouck

1. (Text) Describe the criterion used for declaring CpG islands, explaining why used.
2. What fraction of genes are said to have CpG islands? Can we examine this experimentally by looking at known genes in finished sequence?
3. How many CpG islands are there? How do they correlate with CpG content (of surrounding DNA)? With chromosome? With gene content? With house keeping genes?

4.3 Correlating the Sequence with Cytogenetics — Ewan Birney

1. Where do we have cytogenetic landmarks—i.e., spots in the sequence that have been assigned cytogenetic locations? How can we reflect this information on the map?
2. (Figures) Tie points of the GP to the cytogenetic map. This should be separate from the density plots! Arek Kaspryck at Sanger has nice plotting software

Bruce Birren and Barb Trask have done this
Kirsch and Reed at NCI have done FISH mapping.
Sanger Center has data

We want a paragraph about the degree of cytogenetic mapping and some pretty pictures of nicest chromosomes.

3. How can we correlate CG content with chromosome banding pattern? This requires much greater precision.

4.4 SNPs — Eric Lander (to coordinate)

- ESL to discuss coordination with Bentley/Altshuler

4.5 Caveats — Ewan Birney

(Text) Ewan will describe the many caveats that must be attached to our analyses

NHGRI ROUTE SLIP

Please Circulate

Division of Extramural Research	Dr. Mark Guyer	------	Message
	~~Dr. Bettie Graham~~	✓	FYI
	~~Dr. Elise Feingold~~	✓	
	~~Dr. Jane Peterson~~	✓	
	~~Dr. David Benton~~	✓	
	~~Dr. Jeff Schloss~~	✓	
	~~Dr. Eric Meslin~~	✓	
	~~Ms. Elizabeth Thomson~~	✓	
	~~Ms. Joy Boyer~~	✓	
	Ms. Anita Allen	------	
	Ms. Peggy Whittington	------	
	Ms. Charlotte Quinn	------	
	Ms. Stephanie Reeves-Walker	------	
Office of Scientific Review	~~Dr. Ken Nakamura~~	✓	
	~~Dr. Rudy Pozzatti~~	✓	
	Ms. Gwendolyn Williams	------	
	~~Dr. Jerry Roberts~~	✓	
Grants Managment	Ms. Jean Cahill	------	
	Ms. Sally York	------	
	Ms. Linda Hall	------	
	Ms. Diane Patterson	------	
	Ms. Tara Mowery	------	
	Ms. Monika Yakovich	------	
Office of Information Systems Management	Ms. Carol Martin	------	
			Date: 3/5/97
			Return to: Anita
			From: Mark

Thursday 27th February 1997

2015 Registration

2030 Cocktail Party
Adams Room, Mezzanine Floor

Friday 28th February 1997

0700 Breakfast at Leisure - *Tiara Room, Mezzanine Floor*

Catherine & Victoria , Conference Suite

0830 CHAIRMAN'S INTRODUCTION: MICHAEL MORGAN

0830-1300 ***Session I PROGRESS, STRATEGIES AND DEVELOPMENTS***
Progress reports from each sequencing group.
Speakers are asked to address the following:
a) Effectiveness of strategies for
i) Construction of sequence ready maps
ii) producing finished sequence (finished meaning the quality that the group is Willing to submit to a public database as finished)
b) Incorporation of new libraries into production lines - how will this be achieved?
c) Other bottlenecks, in particular, plans for addressing the finishing of sequence data.
d) Brief report on new technologies

CHAIR OF SESSION: DAVID COX

0835 John Sulston
0850 ✓Robert Waterston
0905 ✓Thomas Hudson
0920 ✓Craig Venter
0935 Richard Gibbs
0950 David Cox
1005 Fiona Francis
1020 ✓Jean Weissenbach
1035 ✓John Mattick
1050 ✓Andre Rosenthal

1105 Morning Coffee - *Lobby Area*

1125 ✓Phil Green
1140 ✓Ellson Chen
1155 Yoshiyuki Sakaki

1210 ✓Asao Fujiyama
1225 ✓Glen Evans
1240 ✓Michael Palazzolo
1255 ✓Bruce Roe

1310 **Luncheon** - ***Tiara Room, Mezzanine Floor***

1415 ***Session II SEQUENCING QUALITY AND COSTS***

CHAIRMAN: FRANCIS COLLINS

Round Table Discussion

Aims of this session are to discuss:
Sequence quality standards:
Should a universal standard addressing base accuracy, coverage and number of gaps per Mb be adopted?
Can a standard/uniform way of measuring the cost of producing sequence be agreed upon?

1600 **Afternoon Tea** - ***Lobby Area***

Session II continues: DATA RELEASE

CHAIRMAN: FRANCIS COLLINS

Round Table Discussion
Aims of this session are to discuss:
How have different groups implemented the conclusions from last years meeting?
Should these conclusions be revisited?
How can the usefulness of very rapid release be assessed?

1800 Close of Session

1930 **Pre Dinner Drinks** - ***Harbourfront Restaurant, Front Street***

2000 **Conference Dinner** - ***Harbourfront Restaurant, Front Street***

Saturday, 1st March 1997

0700 **Breakfast at Leisure** - ***Tiara Room,Mezzanine Floor***

Catherine & Victoria , Conference Suite

09.00-1230 ***Session III ALLOCATION OF REGIONS/ETIQUETTE FOR SHARING***

CHAIRMEN: JOHN SULSTON AND ROBERT WATERSTON

Round Table Discussion

Aims of this session are to discuss mechanisms for the allocation of genomic regions for sequencing:
Territorial Claims- How much sequence is it appropriate to stake out;
role of HUGO and local Web sites
What will happen when more than one group is interested in sequencing a particular region?
What will happen when a group does not meet its commitment to complete a particular region?

1100 **Morning Coffee** - ***Lobby Area***

Session continues in meeting room

12.30 **Luncheon** - ***Tiara Room, Mezzanine Floor***

1400-1800 ***Session IV INTERPRETATION***

CHAIRMAN: DAVID BENTLEY

Round Table Discussion

The aims of this session are:
a) Annotation standards: What level of annotation is appropriate for large-scale genomic sequencing laboratories?
b) EST sequencing/full length cDNA sequencing: What role can such sequences play in assembling and interpreting genomic sequence?
c) Mouse Sequencing: What role can it playin interpreting human sequence?
How much sequence is required to assess its value?
What strategies should be investigated?

1600

Afternoon Tea - ***Lobby Area***

1630 *Session V FUTURE MEETINGS:*

CHAIRMAN: MICHAEL MORGAN

Round Table Discussion

The aims of this session are to discuss whether this meeting should be held next year? or beyond next year?

1800 Close of Session

1930 Pre dinner drinks - *Bermuda Room Foyer*

2000 Dinner - *Bermuda Room, Mezzanine Floor*

Sunday 2nd March 1997

0700 Breakfast at leisure - *Tiara Room, Mezzanine Floor*

Delegates depart at leisure during the day

Dr Mark Adams
The Institute for Genomic Research
9715 Medical Center Drive
Rockville MD 20850
USA

[redacted]

Dr David Bentley
The Sanger Centre
Wellcome Trust Genome Campus
Hinxton
Cambridge CB10 1SA
UK

[redacted]

Dr Elbert Branscomb
Lawrence Livermore National Laboratory
7000 East Avenue
L-452 Livermore
CA 94550
USA

[redacted]

Dr Graham Cameron
The European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton
Cambridge
CB10 1SD
UK

☎
Fax: [redacted]

Dr Ellson Chen
ACGT
ABD-PE
850 Lincoln Center Drive
Foster City CA 94404
USA

[redacted]

Dr Francis Collins
National Institutes of Health
National Human Genome Research
Institute
31 Center Drive MSC 2152
Building 31, Room 4B09
Bethesda MD 20892-2152
USA

[redacted]

Professor David Cox
Stanford University School of Medicine
Department of Genetics
Room M336
Stanford CA 94305
USA

[redacted]

Dr Pieter de Jong
Rosewell Park Cancer Institute
Human Genetics Department
Elm & Carlton Streets
Buffalo NY 14263
USA

[redacted]

Dr Glen Evans
University of Texas Southwestern
Medical Center
Human Growth & Development
6000 Harry Hines Boulevard
Dallas TX 75235-8591
USA

[redacted]

Professor Asao Fujiyama
National Institute of Genetics
Division of Human Genetics
1111 Yata
Mishima
Shizuoka 411
Japan

[redacted]

Dr Fiona Francis
Max-Planck-Institut Für Molekulare
Genetik
Ihnestrasse 73
Berlin 14195
Germany

[redacted]

Dr Richard Gibbs
Department of Molecular and Human
Genetics
Baylor College of Medicine
One Baylor Plaza
BCMS-930
Houston TX 77030
USA

[redacted]

Dr Eric Green
National Institutes of Health
National Human Genome Institute
49 Convent Drive
Building 9, Room 2A08
Bethesda MD 20892-4431
USA

[redacted]

Dr Phil Green
University of Washington
Department of Molecular Biotechnology
GJ10
4909 25th Avenue NE
Seattle WA 98195
USA

[redacted]

Dr Mark Guyer
National Institutes of Health
National Human Genome Research
Institute
38 Library Drive
MSC 6050
38A/604
Bethesda MD 20892-6050
USA

[redacted]

Dr Trevor Hawkins
Whitehead Institute/MIT Center for
Genome Research
One Kendall Square
Building 300
Cambridge MA 02139-1561
USA

[redacted]

Dr LaDeana Hillier
Washington University School of Medicine
Genome Sequencing Center
Box 8501
4444 Forest Park Avenue
St Louis MD 631
USA

[redacted]

Dr Thomas Hudson
Whitehead Institute/MIT Center for Genome Research
One Kendall Square
Building 300
Cambridge MA 02139
USA

[redacted]

Dr Ursula Hurtenbach
DLR - Projekttrdger des BMBF
Sudstrasse 125
D-53175 Bonn
Germany

[redacted]

Dr Elke Jordan
National Institutes of Health
National Center for Human Genome Research
31 Center Drive - MSC 2152
Building 31, Room 4B09
Bethesda MD 20892-2152
USA

[redacted]

Dr Jill Kent
The Wellcome Trust
183 Euston Road
London
NW1 2BE

[redacted]

Dr Ung-Jin Kim
California Institute of Technology
Division of Biology 147-75
Pasadena CA 91125
USA

[redacted]

Dr David Lipman
National Center for Biotechnology Information
National Institutes of Health
National Library of Medicine
Building 38A, 8N805
8600 Rockville Pike
Bethesda MD 20894
USA

[redacted]

Dr Richard McCombie
Cold Spring Harbor Laboratory
P.O.Box 100
Cold Spring Harbor
NY 11724
USA

[redacted]

Professor John Mattick
University of Queensland
Centre for Molecular and Cellular Biology
St Lucia
Brisbane Queensland
Australia 4072

[redacted]

Dr Catherine Moody
Medical Research Council
20 Park Crescent
London
W1N 4AL
UK

[redacted]

Dr Michael Morgan
The Wellcome Trust
183 Euston Road
London
NW1 2BE
UK

[redacted]

Dr Michael Palazzolo
Lawrence Berkeley National Laboratory
One Cyclotron Road
MS 74-157
Berkeley CA 94720
USA

[redacted]

Dr Ari Patrinos
Health Effects & Life Sciences Research
Division
US Department of Energy
19901 Germantown Road
Germantown MD 20874-1290
USA

[redacted]

Dr Jane L. Peterson
National Institutes of Health
National Center for Human Genome
Research
Room 614, Building 38A
Bethesda MD 20892
USA

[redacted]

Professor Bruce Roe
The University of Oklahoma
Chemistry Department
620 Parrington Oval, Room 208
Norman
Oklahoma 73069
USA

[redacted]

Dr Jane Rogers
The Sanger Centre
Wellcome Genome Campus
Hinxton
Cambridge, CB10 1SA
UK

[redacted]

Professor André Rosenthal
Institute of Molecular Biotechnology
Department of Genome Analysis
Beutenbergstrasse 11
D-07745 Jena
Germany

[redacted]

Dr Barbara Skene
The Wellcome Trust
183 Euston Road
London
NW1 2BE
UK

[redacted]

Mrs Jilly Steward
The Wellcome Trust
183 Euston Road
London NW1 2BE
UK

[redacted]

Dr John Sulston
The Sanger Centre
Wellcome Trust Genome Campus
Hinxton
Cambridge CB10 1SA
UK

[redacted]

Professor Gert-Jan Van Ommen
University of Leiden
Department of Human Genetics
P.O.Box 9503
2300 RA Leiden
The Netherlands

[redacted]

Dr Craig Venter
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
USA

[redacted]

Ms Susan Wallace
HUGO Americas
7986D Old Georgetown Road
Bethesda MD 20814
USA

[redacted]

Dr Robert H. Waterston
Washington University School of
Medicine
Genome Sequencing Center
Box 8501
4444 Forest Park Avenue
St Louis MO 631
USA

[redacted]

Professor Sir David Weatherall
Institute of Molecular Medicine
John Radcliffe Hospital
Headington
Oxford
OX3 9DU

[redacted]

Dr Jean Weissenbach
CNRS URA 1922
Genethon
1 rue de l'internationale
BP 60 - 91002 Evry Cedex
France

[redacted]

Action Items

1. Data quality exercise - all grantees
 short-term - one BAC - grantee mtg @ CSH
 longer term - statistical sample of each center's outp

 address - representation to genome?
 clone integrity
 gaps - minimal
 contiguity

2. Cost - audit?
 Who to do it?
 are we involved?
 just advice to grantees?

3. Data release
 how to make a condition of award - by April 1
 on all grants?

4. usefulness of unfinished data
 solicit opinion on web site?

5. find out about Oxford mouse EST, mapping -

1) Data quality — schedule meeting @ CSH.
include all grantees eventually - pilot for pilots

Ask for clear description of q.c.
Internal q.c part of [illegible] rec.

John Spouge - NCBI - MD PhD - could help [illegible]

~~2) Cost~~ Unfinished sequence - → GenBank.

Judging pilots - finished seq - data release
- to clone & assembly funding
accuracy
# gaps
Cost/bp
Contiguity
data release of unfinished seq.

2) Costs — educational model
done now - to help them set up.
ask grantees in [illegible] grp.
[illegible] educational & proposal for rev.
Rev panel - .

3) data release
Send out letter - Summary of Bermuda mtg
- plan to make cond of award -
then put deadline for cond to be in compliance

"Anything DIKG appears ON our Web Page'

Talk w/ Jim re: letter
All grantees -

Funded + unfunded data - Arbitrate opinions?

Annotation - confidence levels reported -
talk to Jim Buck

International contact w/ Ursula re: Genome.

JOHN SULSTON

Summary of targets

Main projects

Work is in progress on the following five chromosomes. Selected regions are the subject of early effort as listed, but further mapping and clone isolation is under way for the majority of each chromosome. See individual project pages for further information.

These regions of Chromosomes 22 and X are being sequenced jointly with GSC, St Louis.

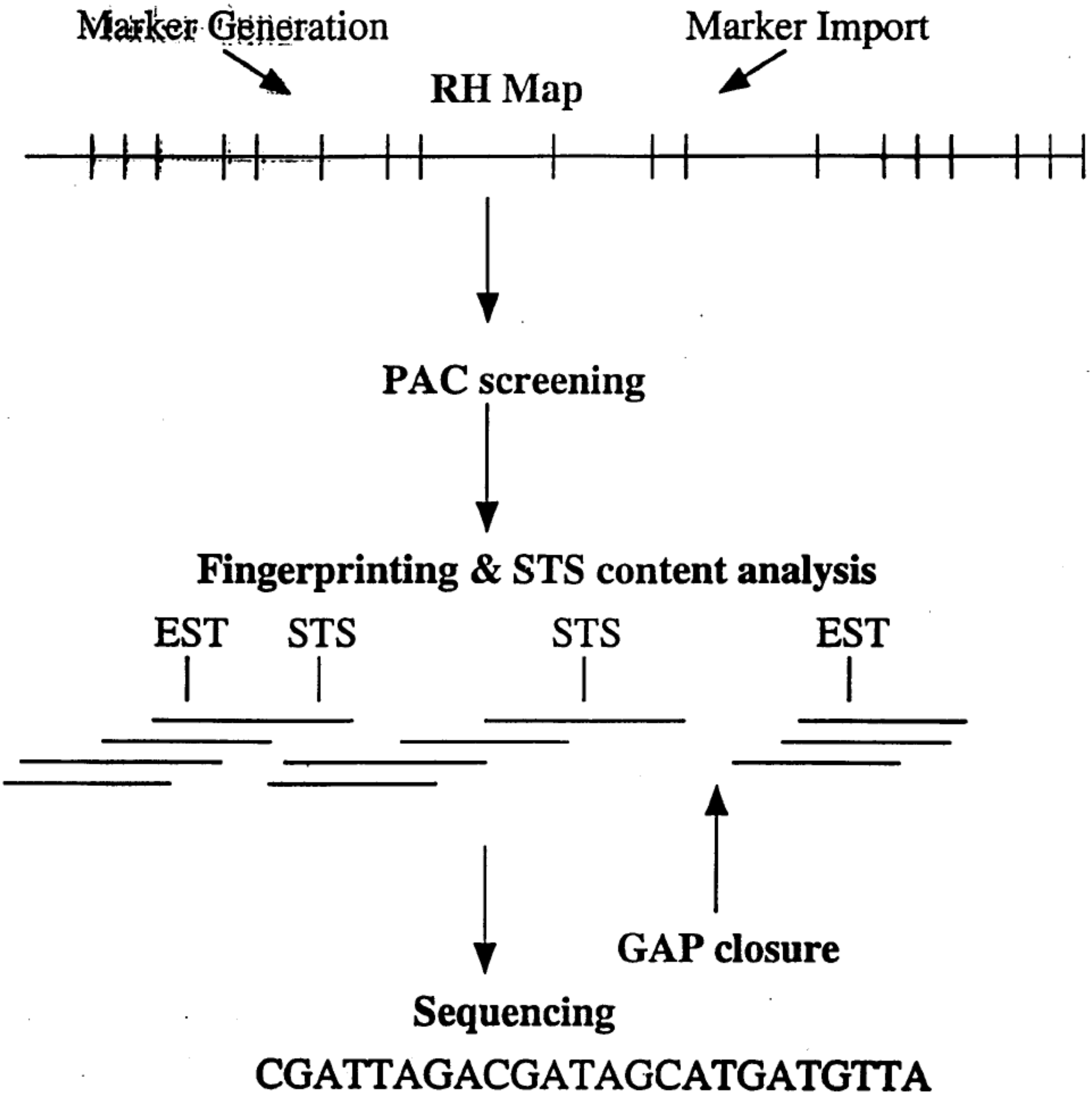
Chromosome 1	300 Mb	1p35-1pter 1pcen-1p13 1q22
Chromosome 6	160 Mb	6p21.3 6p23 6q21 6q27
Chromosome 20	80 Mb	20q11.2-13.1
Chromosome 22	25 of 45 Mb	22q12-13
X chromosome	90 of 150 Mb	Xp Xq22 Xq23-26

Sequencing collaborations

We undertake collaborations to sequence limited regions of specific interest, as listed below:

Chromosome	Size	Region
3p21.3	0.3 Mb	The LUCA6 region
4p	1.6 Mb	The HD region
11p13	0.2 Mb	The PAX6 region
11p15.5	80 kb	
12		The MODY3 region
13q12	0.9 Mb	The BRCA2 region
16p	0.3 Mb	The globin region

See also:


Marker Generation
Marker Import
RH Map
PAC screening
Fingerprinting & STS content analysis
EST
STS
STS
EST
GAP closure
Sequencing
CGATTAGACGATAGCATGATGTTA

Sanger Centre Summary of Human Progress

(all figures are Mb except markers)

Chromosome	1	6	20	22	X	Other	Total
S.C. region	300	160	80	25	90		655
Markers working	3029	2720	1268	951 (+166)	866 (+127)		
[Markers/Mb]	[10.1]	[17.0]	[15.9]	[21.1]	[9.6]		
Coverage in bacterial clones	20	23.9	5.2	19.3	29.0		97.4
Ready for seq	0.8	4.4	1.0	5.0	10.2		17.5
Unfinished seq	0	1.9	0	5.4	4.1	0.5	11.9
Finished seq	0	0.6	0	3.1	7.5	3.4	14.6
Total seq on ftp	0	2.5	0	8.5	11.6	3.9	26.5

Sanger Centre Total Sequence Output (Mb)
February 1997

	Unfinished	Finished	**Total in Public Domain**
Nematode	9.7	29.7	**39.4**
Human	13.8	14.2	**28.0**
Yeasts	0.0	6.2	**6.2**
TB	1.1	2.1	**3.2**
TOTAL	24.6	52.2	**76.8**

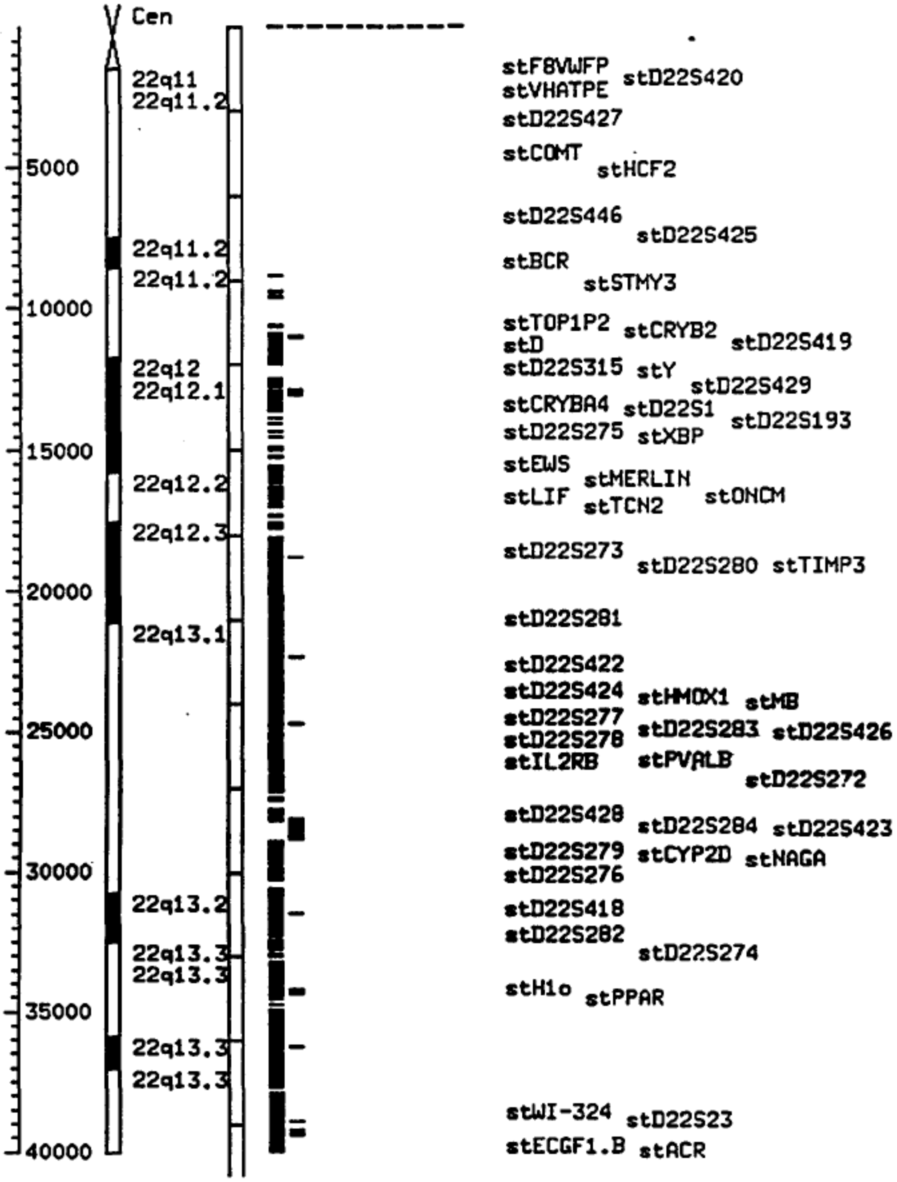
TOTAL FINISHED LAST YEAR 34 Mb

Map Status for Chromosome 22

The picture on the left shows the current status of sequencing for **Chromosome 22.**

Click on the **column of white boxes** to zoom in on an interval of the chromosome. Click on the **red boxes** to see the clones being sequenced. Click on the **marker names** to see a report for that marker (this is still under development).

All sequence data for this region is available from the human sequence directory of our FTP site.



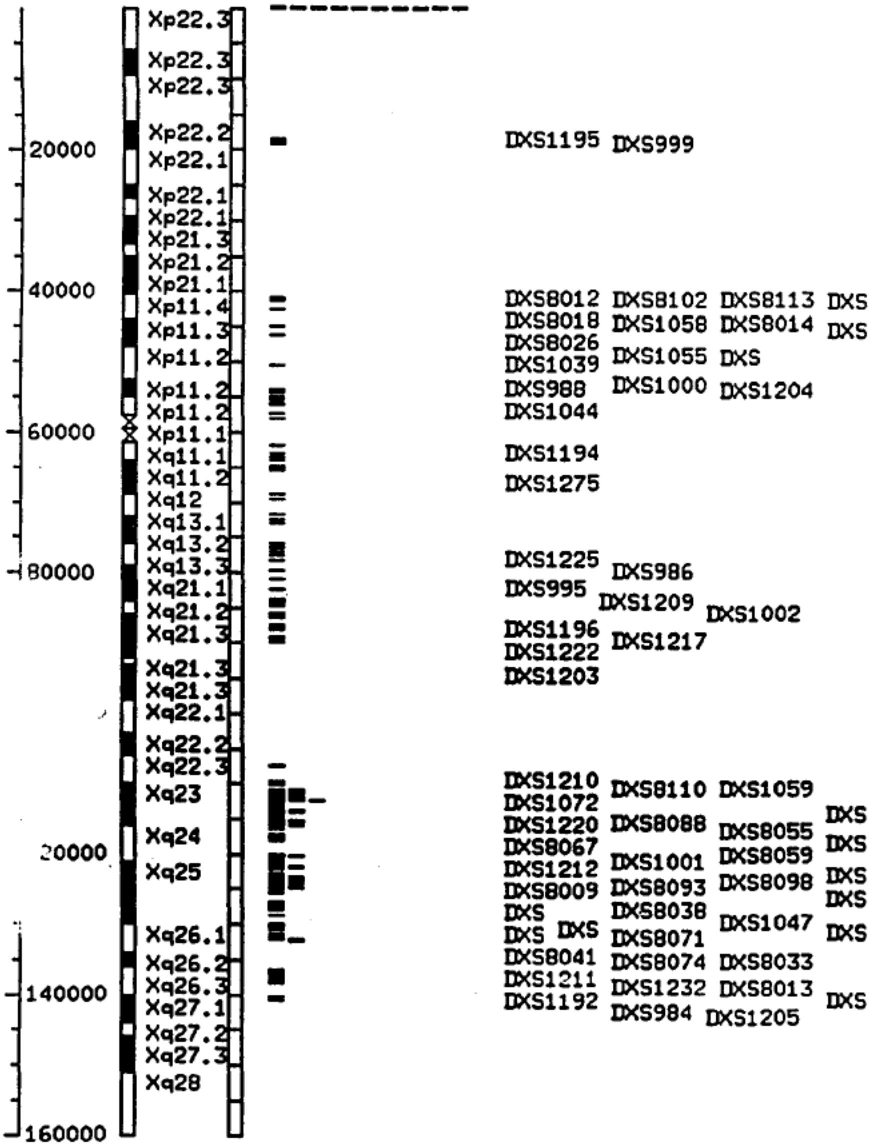
The graphical display was made using acedb

Map Status for Chromosome X

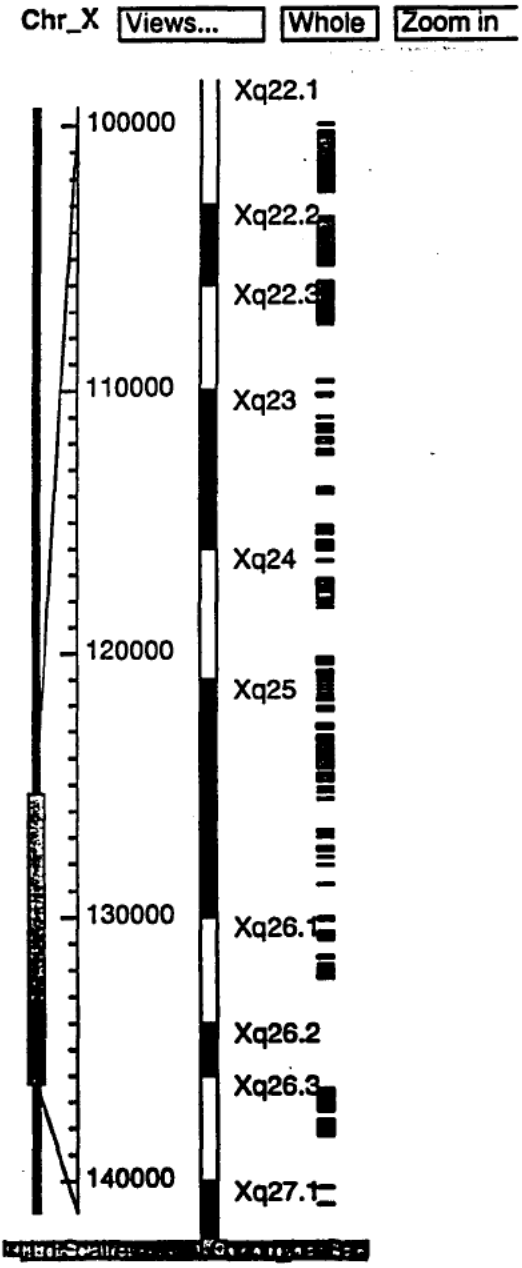
The picture on the left shows the current status of sequencing for **Chromosome X.**

Click on the **column of white boxes** to zoom in on an interval of the chromosome. Click on the **red boxes** to see the clones being sequenced. Click on the **marker names** to see a report for that marker (this is still under development).

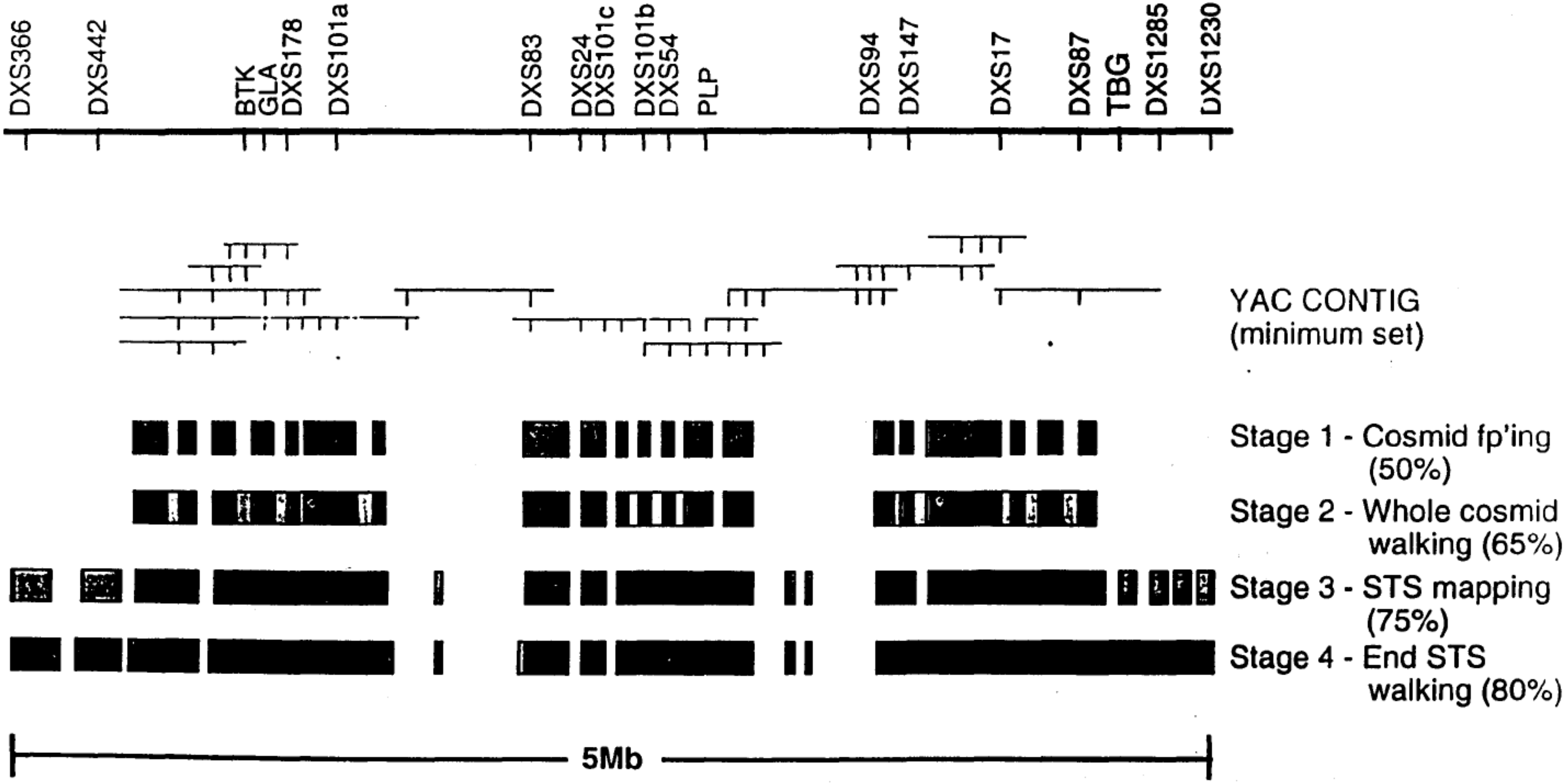
All sequence data for this region is available from the human sequence directory of our FTP site.



The graphical display was made using acedb


Chr_X
Views...
Whole
Zoom in
100000
110000
120000
130000
140000
Xq22.1
Xq22.2
Xq22.3
Xq23
Xq24
Xq25
Xq26.1
Xq26.2
Xq26.3
Xq27.1

Cosmid Coverage In Xq22 From DXS366 To DXS1230



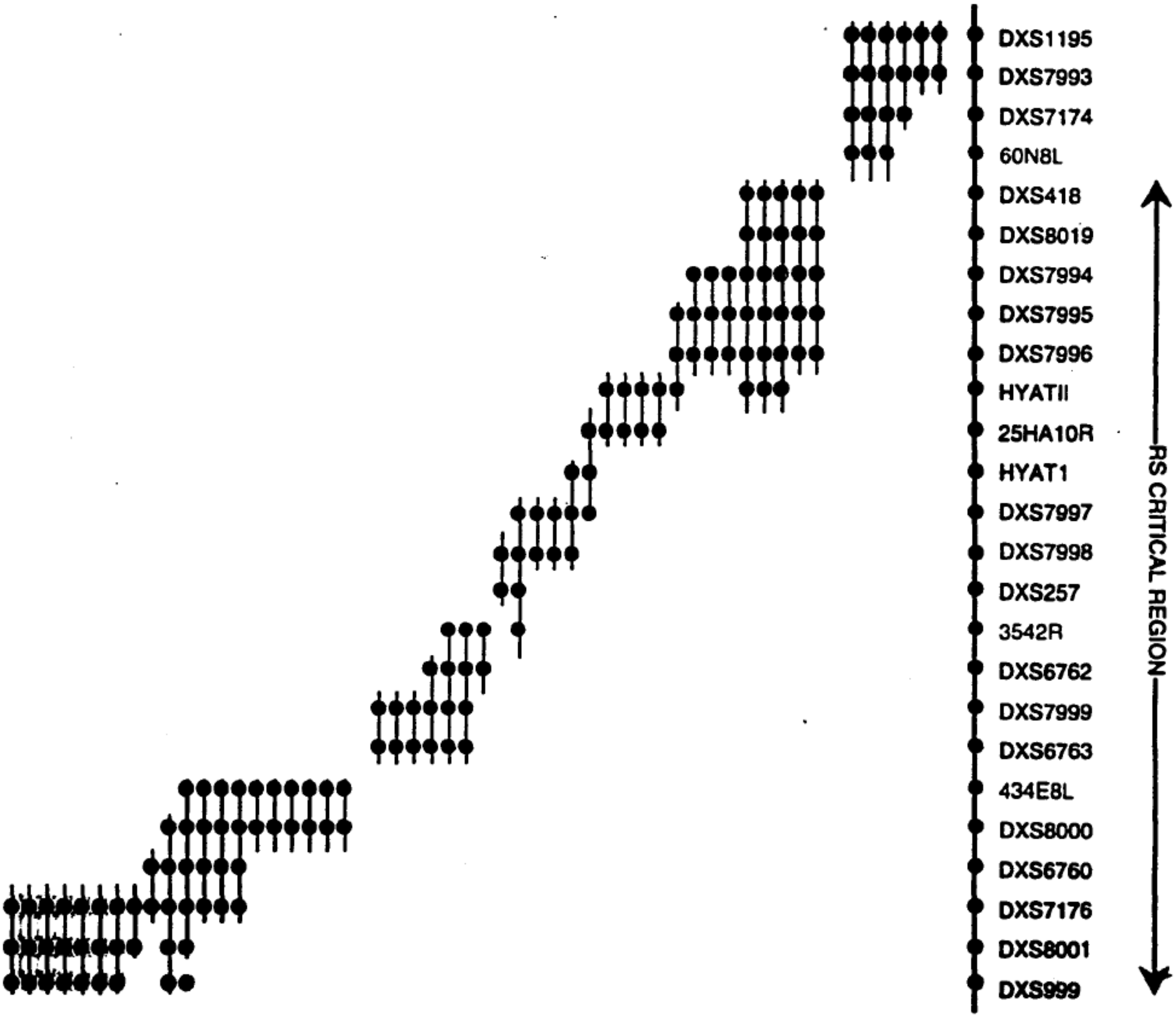
YAC CONTIG
(minimum set)

Stage 1 - Cosmid fp'ing
(50%)

Stage 2 - Whole cosmid
walking (65%)

Stage 3 - STS mapping
(75%)

Stage 4 - End STS
walking (80%)


DXS1195
DXS7993
DXS7174
60N8L
DXS418
DXS8019
DXS7994
DXS7995
DXS7996
HYATII
25HA10R
HYAT1
DXS7997
DXS7998
DXS257
3542R
DXS6762
DXS7999
DXS6763
434E8L
DXS8000
DXS6760
DXS7176
DXS8001
DXS999
RS CRITICAL REGION

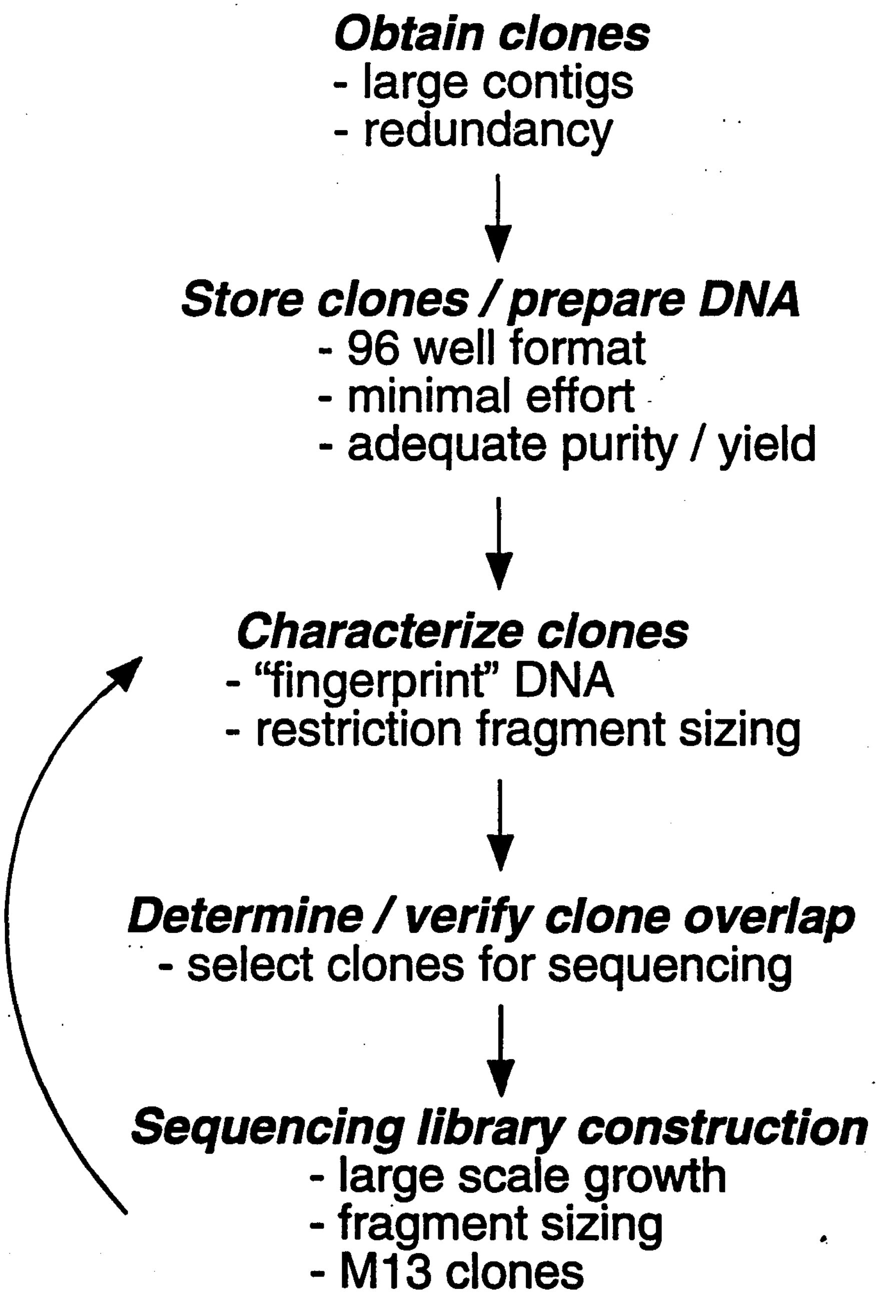
BOB
WATERSTON

970226

in libcore	3.50 Mb
in Shotgun	15.1 Mb
in Finishing	11.6 Mb
Finished	1.85 Mb
Submitted	2.95 Mb

7, 22, X

RHW



Obtain clones
- large contigs
- redundancy

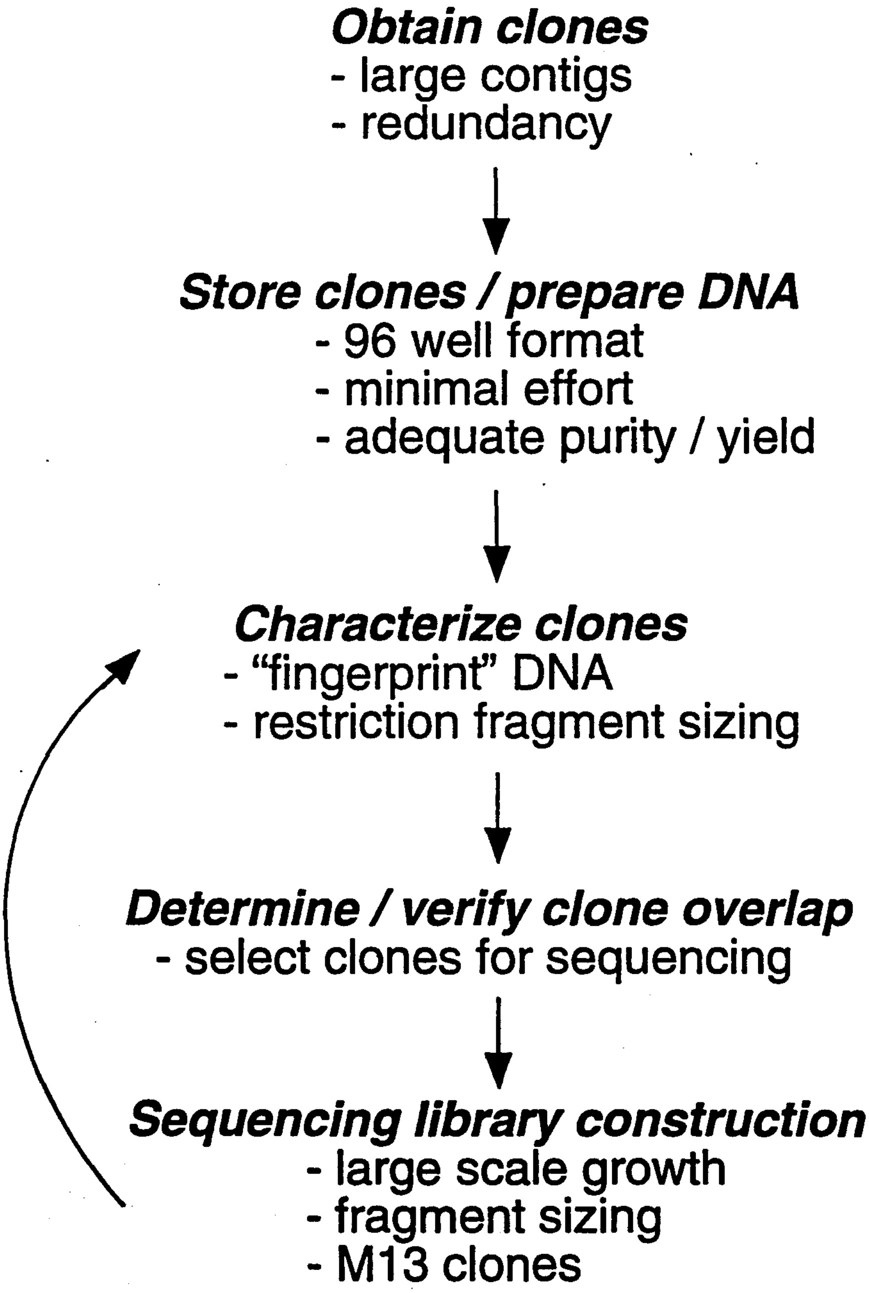
Store clones / prepare DNA
- 96 well format
- minimal effort
- adequate purity / yield

Characterize clones
- "fingerprint" DNA
- restriction fragment sizing

Determine / verify clone overlap
- select clones for sequencing

Sequencing library construction
- large scale growth
- fragment sizing
- M13 clones

RHW



Obtain clones
- large contigs
- redundancy

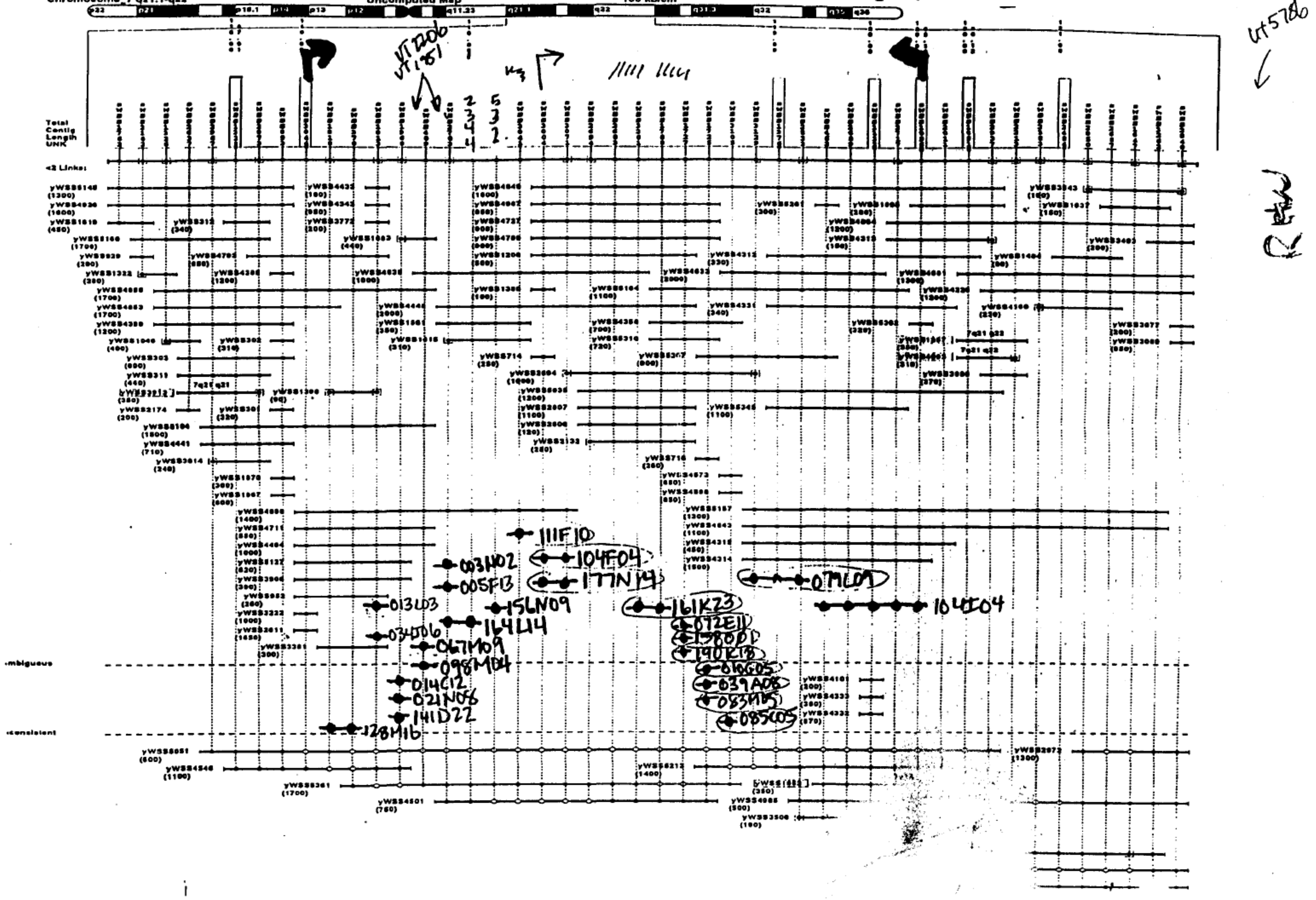
Store clones / prepare DNA
- 96 well format
- minimal effort
- adequate purity / yield

Characterize clones
- "fingerprint" DNA
- restriction fragment sizing

Determine / verify clone overlap
- select clones for sequencing

Sequencing library construction
- large scale growth
- fragment sizing
- M13 clones

sWSS370
Segmap V. 3.45 Data File Date: Thu Nov 16 12:21:19 1995
Chromosome_7 q21.1-q22

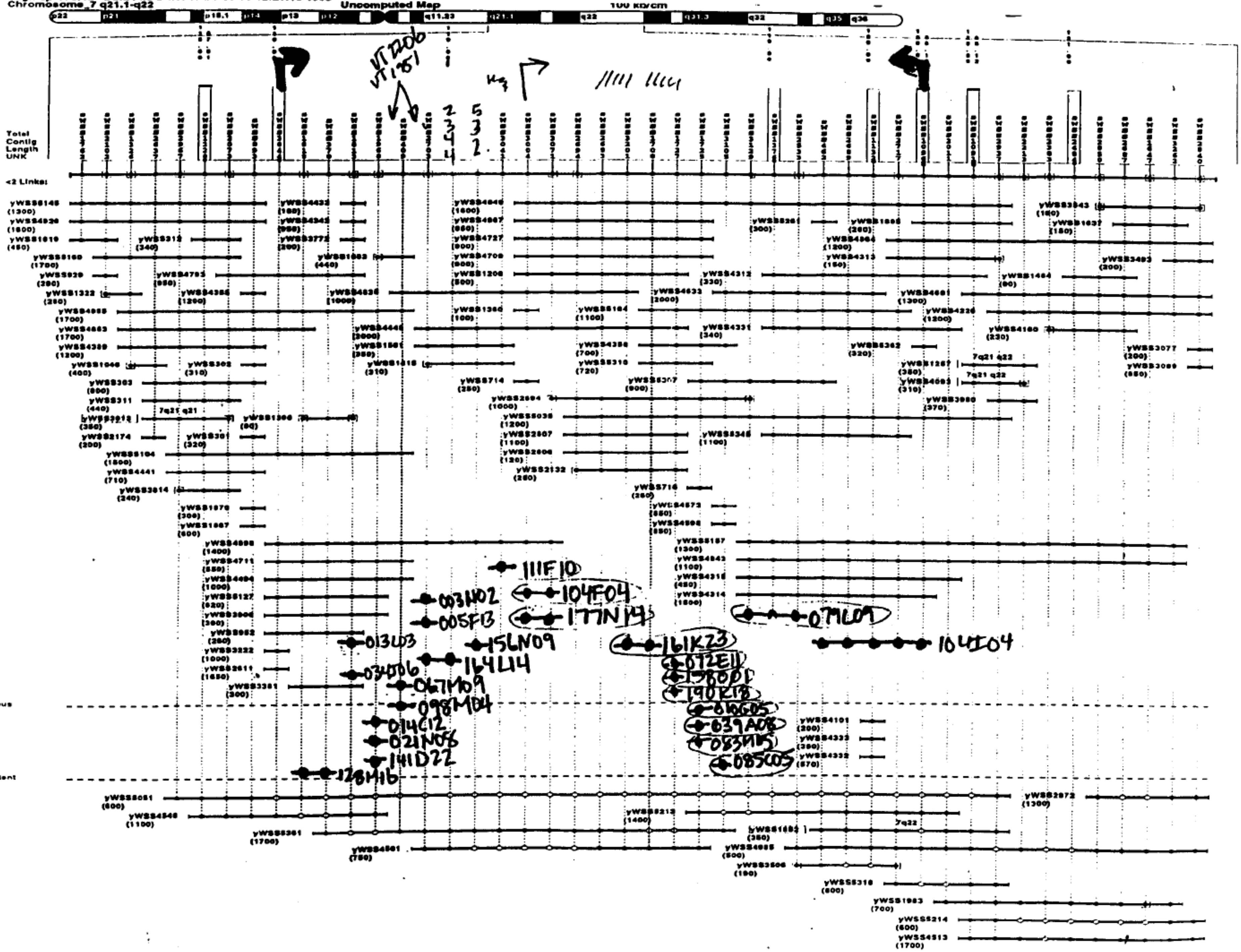


CCM1 Region

sWSS370
Segmap V. 3.45 Data File Date: Thu Nov 16 12:21:19 1995
Chromosome_7 q21.1-q22

Uncomputed Map

100 kb/cm

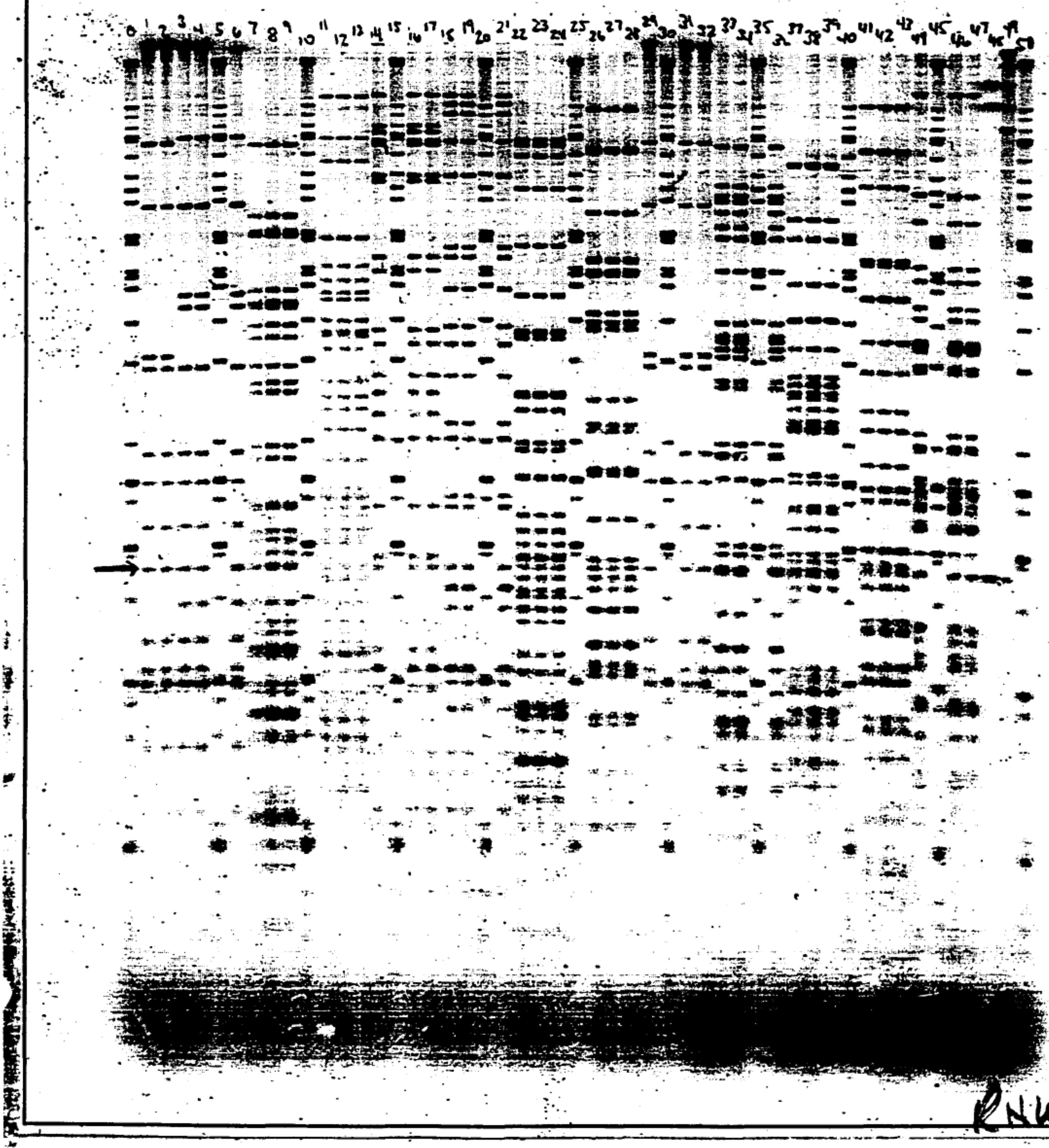


~~e:\m~~m61107h.ds , Range = 100.3-28503 rfu, 105%

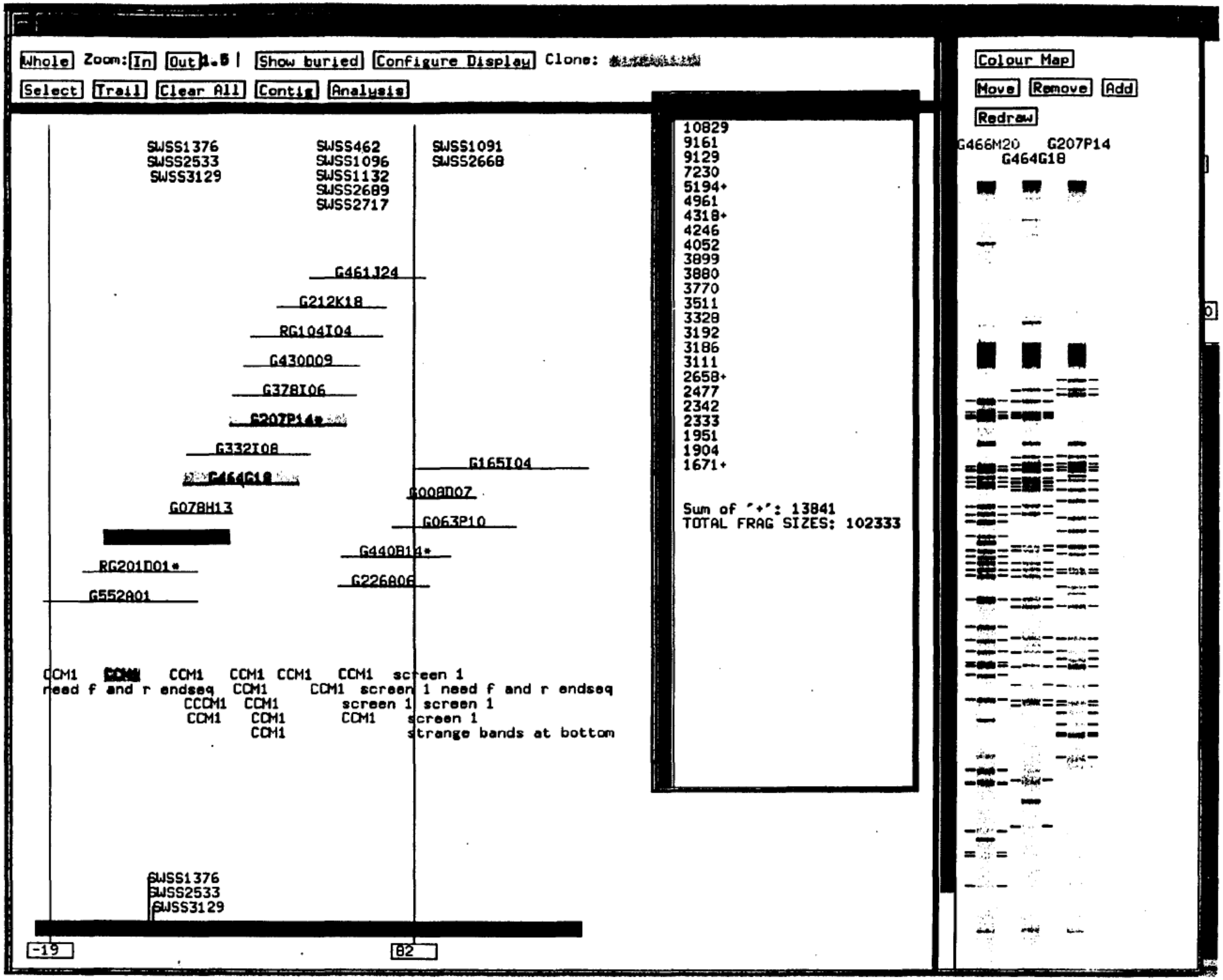
A_013 (A2-D5)

gel26

①



RNW



```
SWSS1376        SWSS462     SWSS1091
SWSS2533        SWSS1096    SWSS2668
 SWSS3129       SWSS1132
                SWSS2689
                SWSS2717

G461J24
G212K18
RG104I04
G430O09
G378I06
G207P14*
G332I08
G165I04
G464G18
G008D07
G078H13
G063P10
[redacted]
G440B14*
RG201D01*
G226A06
G552A01

CCM1    CCM1    CCM1    CCM1  CCM1    CCM1   screen 1
need f and r endseq   CCM1       CCM1  screen 1 need f and r endseq
                CCCM1   CCM1          screen 1 screen 1
                 CCM1    CCM1         CCM1    screen 1
                         CCM1                 strange bands at bottom

SWSS1376
SWSS2533
 SWSS3129
```

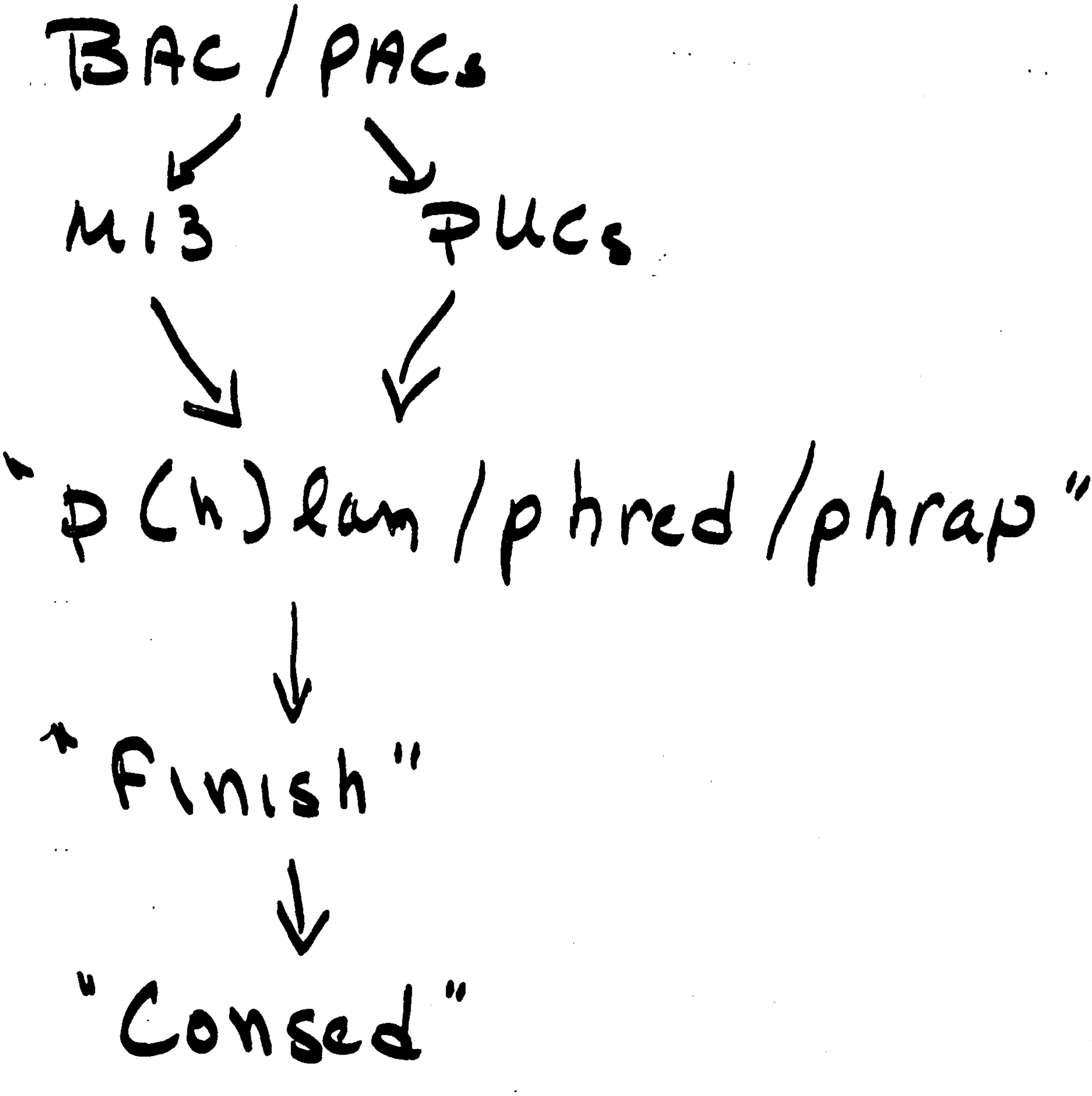
```
10829
9161
9129
7230
5194+
4961
4318+
4246
4052
3899
3880
3770
3511
3328
3192
3186
3111
2658+
2477
2342
2333
1951
1904
1671+

Sum of "+": 13841
TOTAL FRAG SIZES: 102333
```

RHW

Producing Sequence

Shotgun / directed



BAC / PACs

M13

PUCs

"p(h)lam / phred / phrap"

* "Finish"

"Consed"

RHW

Software for human decision making.

data tracking -

central database

bar coding

getlanes / plan / phred / phrap

finish - rearraying

RHW

Technologies

Present -

64-72 lanes on 377

gel loaders

Amersham dye terminators

Transposons

Future

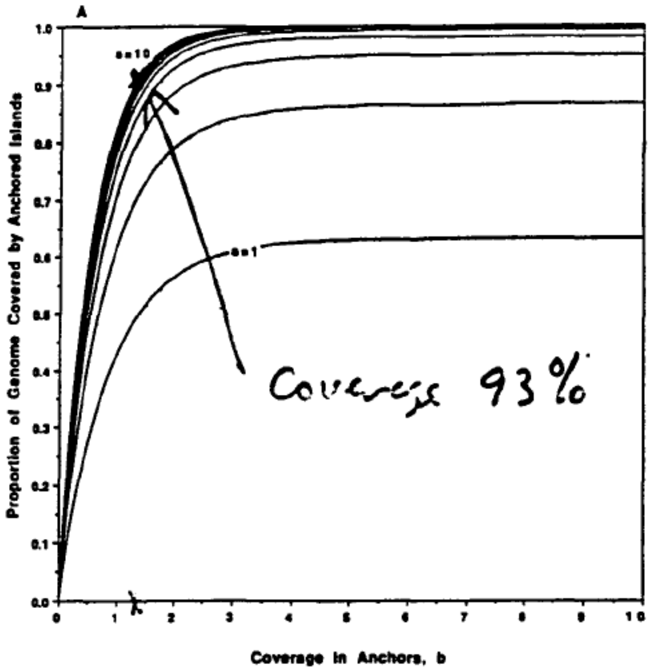
96 lanes on 373, 377

pipetting station

U.W. sequencer

RHW

WHITEHEAD

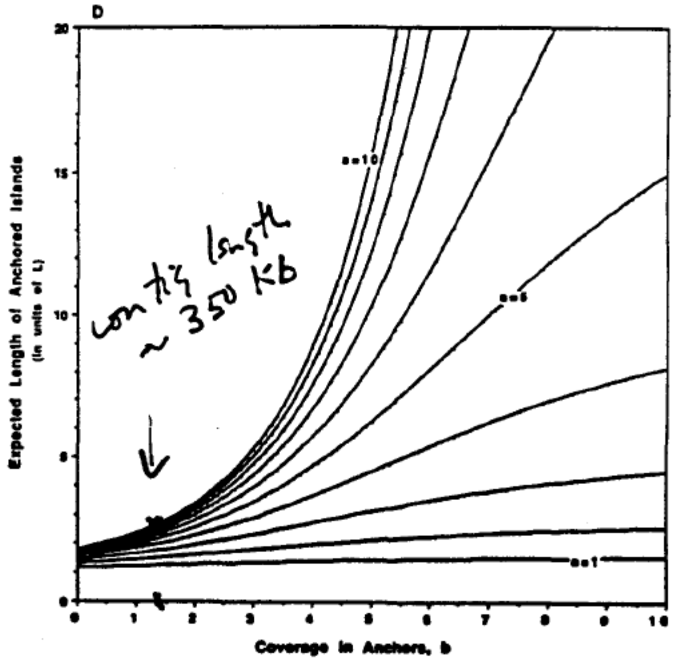


Coverage 93%

ASSUME - 1STS/100 KB.

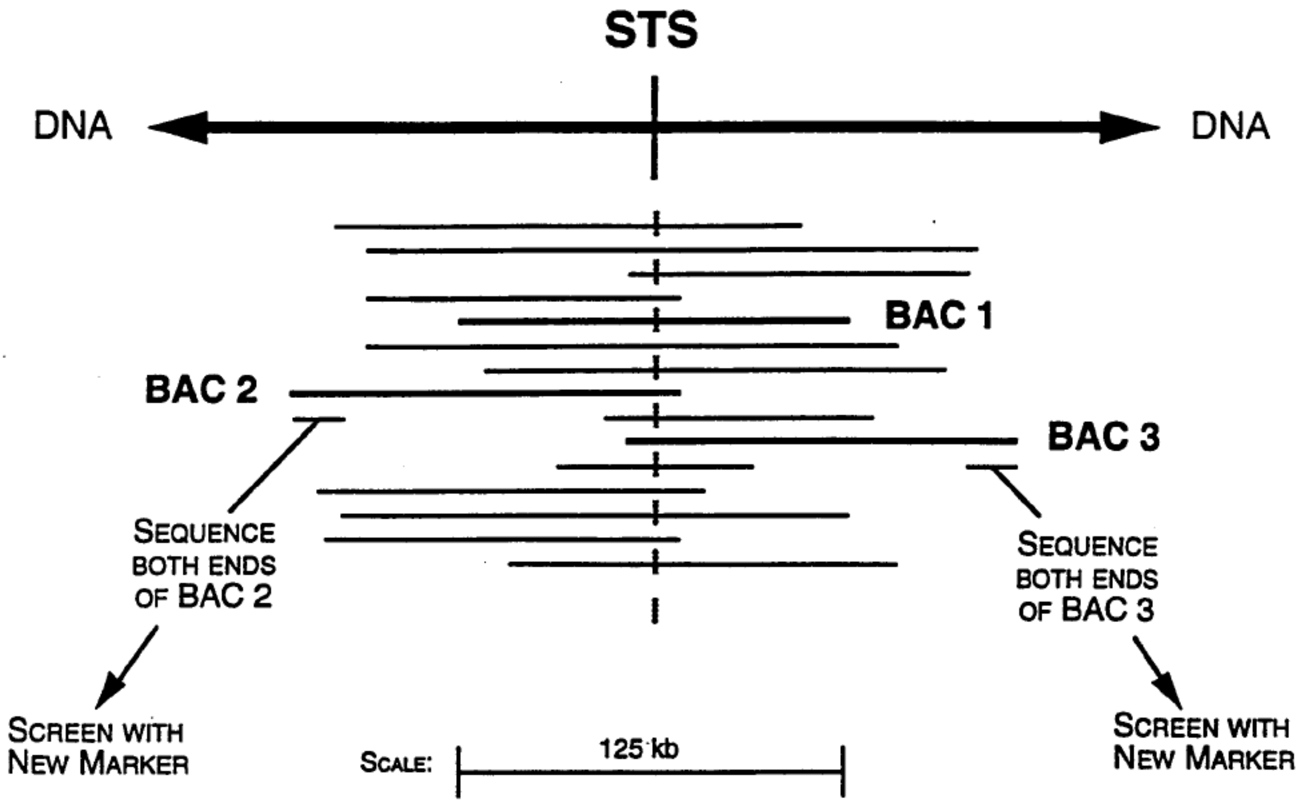
ASSUME - 10X LIBRARY

ASSUME - CLONE ~ 140 KB

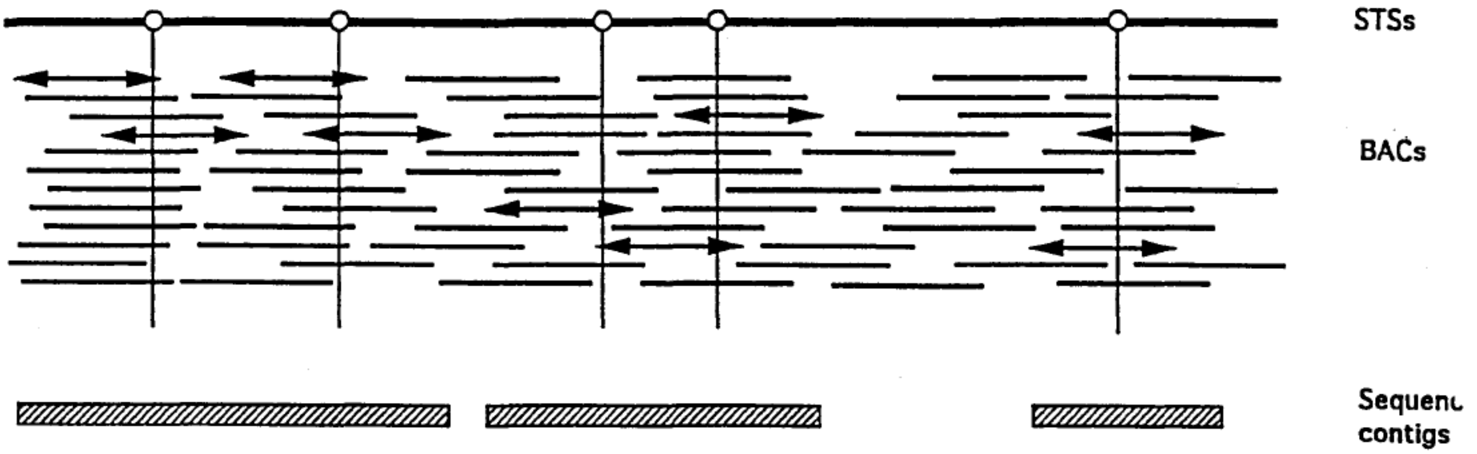


contig length ~ 350 Kb

HUDSON

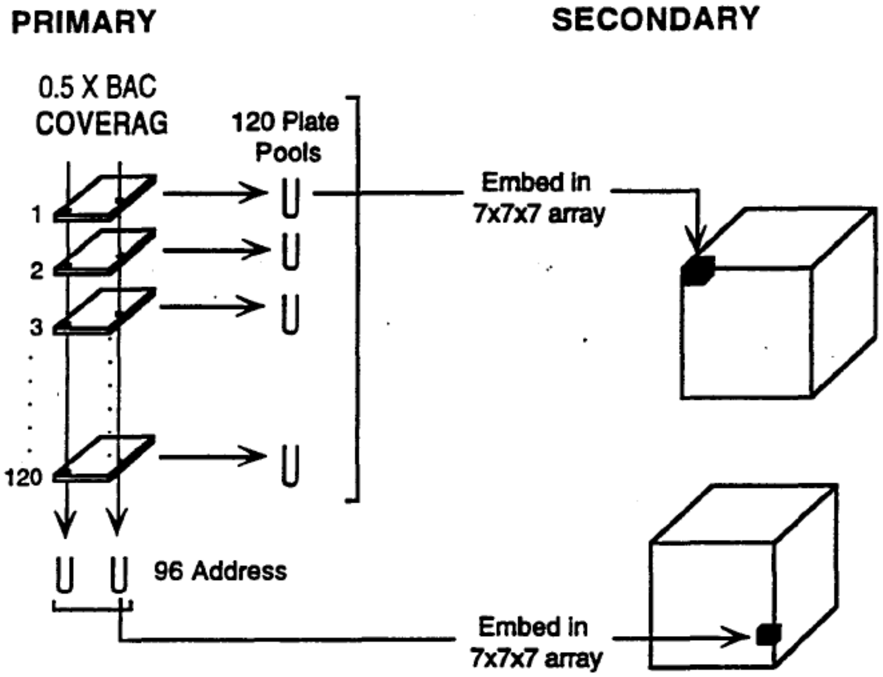

STS
DNA
DNA
BAC 1
BAC 2
BAC 3
Sequence both ends of BAC 2
Sequence both ends of BAC 3
Screen with New Marker
Screen with New Marker
Scale:
125 kb

Hudson

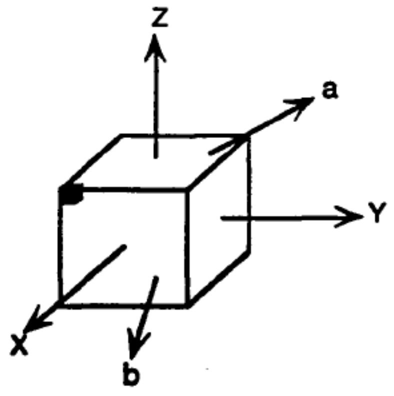


Hudson

1. BAC POOLING SCHEME



2. SCREENING BAC POOLS



70 PCR Assays For a 0.5X Library

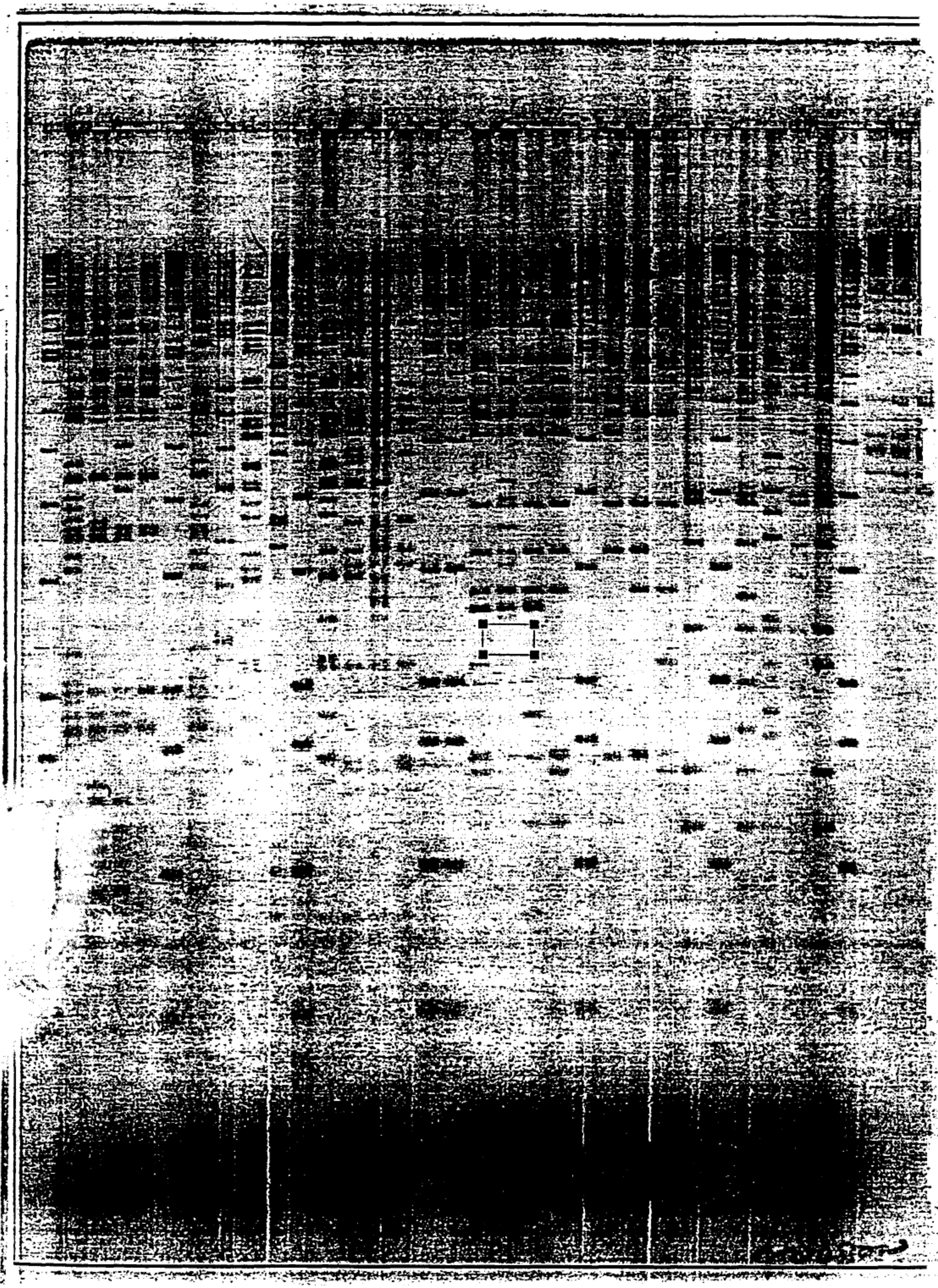
3. SCREENING 20X BAC LIBRARIES

2800 PCR ASSAYS for a 20X Library

GENOMATRON:	**300,000 PCRs/day**
CAPACITY:	**100 STSs screened/day**

Hudson

C:\I ATA\adedig10.ds , Range = 0.000-100004 rfu, 106%



Hawkins

Finishing Focus

- Lab/Automation
  - Biochemical 'tool box' of methods
  - Learning Process
  - Finishatron automation
- Computer
  - Automated workflow
    - TaskMaster LIMS
    - Trout signal processing/base calling
    - Alewife assembler
    - Autoeditor
    - List generator
  - Post Sequencing Varification
    - Big Brother
    - Restriction enzyme/forward-reverse path checking

HAWKINS

Production Finishing

- Finishing should be a production line
  - 80-90% of clones must be treated within the system for optimal throughput
  - Set-up 'swat team' for completion of more unusual clones
- Finishing by Numbers
  - Set of methods and landmarks for progress and automation streamlining

Hawkins

Whitehead Institute/MIT Genome Sequencing Project

View by Progress *Last updated January 31st 1997*

Total Finished	2175 Kb
Total In Finishing	659 Kb
Total	2834 Kb

Clone name	Internal Name	Clone Type	Size (kb)	Location	Status	Gaps	Completed
L196C8	L3	Cosmid	39	Human9q34	Finished	0	Sequence
L2C9F1	L5	Cosmid	39	Human9q34	Finished	0	Sequence
S30E11	L6	Cosmid	38	Human9q34	Finished	0	Sequence
L124D6	L15	Cosmid	40	Human9q34	Finished	0	Sequence
S272C1	L16	Cosmid	33	Human9q34	Finished	0	Sequence
S63C9	L19	Cosmid	40	Human Y	Finished	0	Sequence
5195	L22	P1	79	Mouse 19	Finished	0	Sequence
B287E5	L24	BAC	140	Mouse 9	Finished	0	Sequence
1204	L36	Cosmid	42	Mouse 11	Finished	0	Sequence
46A6	L43	Cosmid	44	Human Y	Finished	0	Sequence
L101D11	L27	Cosmid	46	Human9q34	Finished	0	Sequence
-	L18	Cosmid	29	Mouse 11	Finished	0	Sequence
182E3	L8	Cosmid	46	Human9q34	Finished	0	Sequence
152F5	L10	Cosmid	49	Human9q34	Finished	0	Sequence
44J6	L107	BAC	136	Human 17	Finished	0	Sequence
OC401	L53	PAC	107	Human13	Finished	0	Sequence
320L17	L26	BAC	146	Mouse 9	Finished	0	Sequence

HAWKINS

The Learning Curve

April '96

- Infrastructure
  - Computing
  - Team of 20 people, 6 ABIs
  - Team Leaders
- Development
  - Procedures that scale
  - New electrophoresis conditions/devices
- Library Construction
  - Skills/early QC
- Production Sequencing
  - Sequatron Systems
  - WorkFlow
- QC/QA
  - Reagents
  - Gel to gel
  - Projects
  - Auto trend detection

Dec '96

HAWKINS

Current Issues

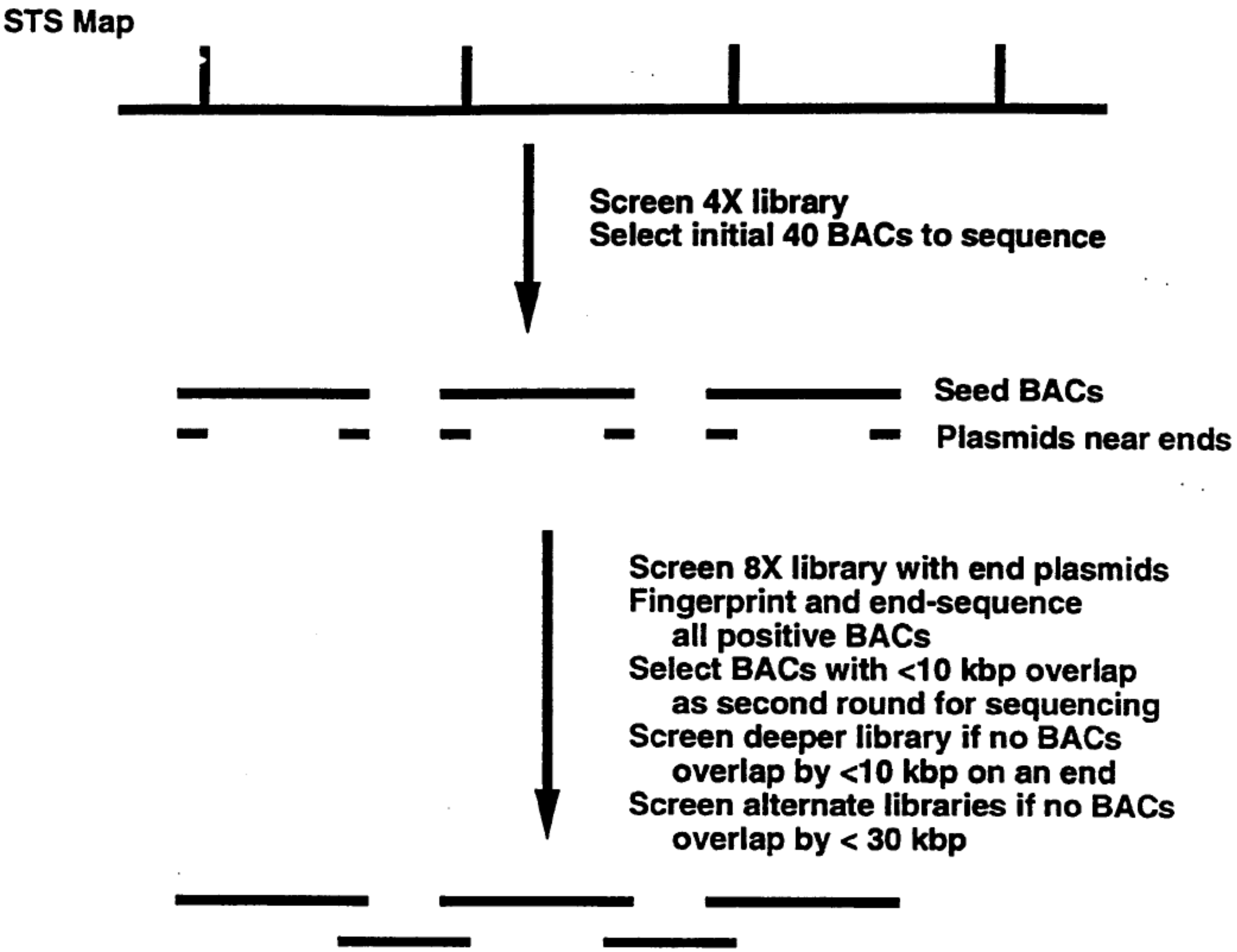
- Finishing
  - Lab issues
  - Computer issues
  - Automation
- Interpretation
  - Human-Mouse synteny
  - Computational methods
    - Ken Fasman

Current

Near Future

MARK ADAMS

TIGR/CalTech Mapping Strategy



STS Map

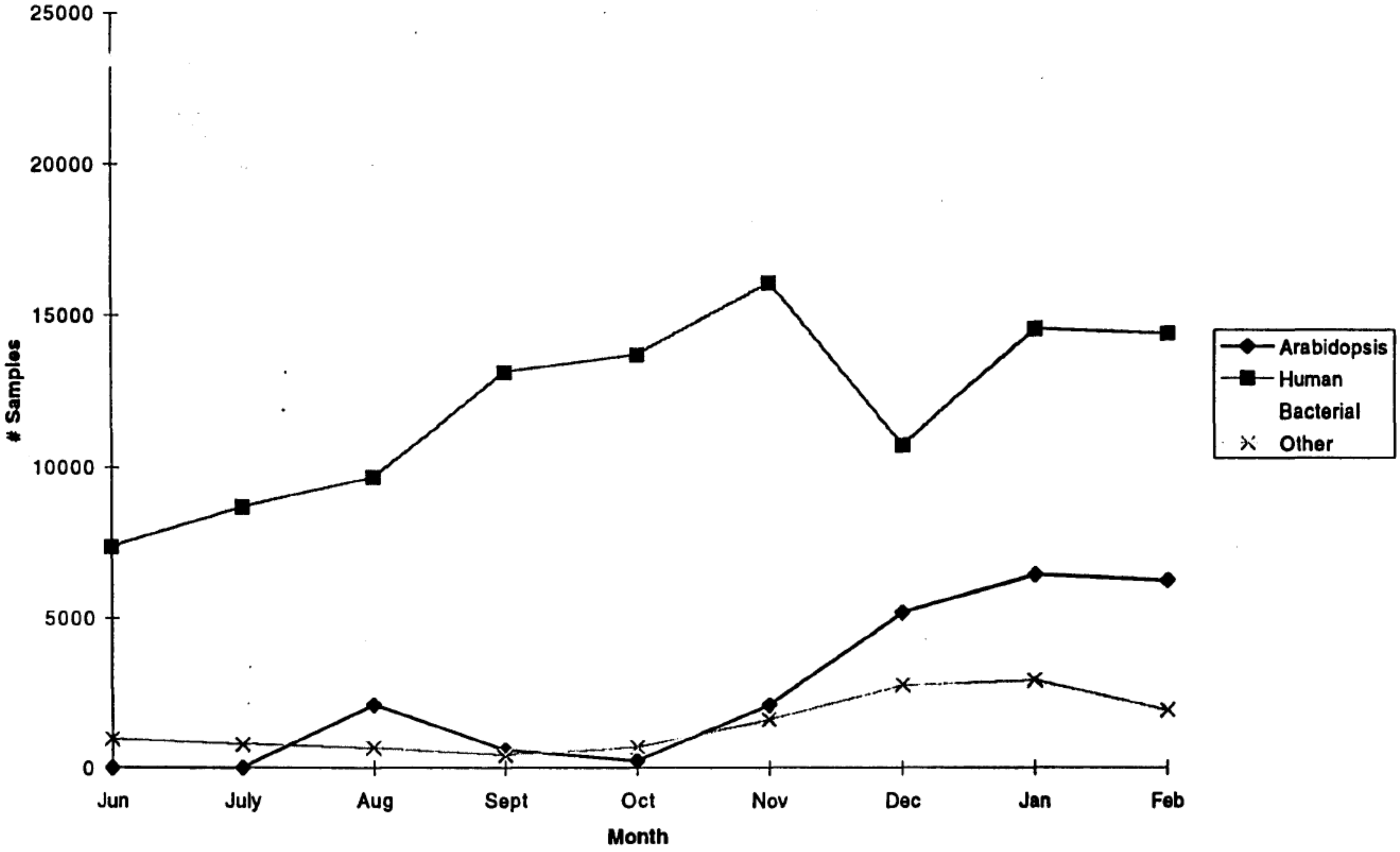
Screen 4X library
Select initial 40 BACs to sequence

Seed BACs

Plasmids near ends

Screen 8X library with end plasmids
Fingerprint and end-sequence
all positive BACs
Select BACs with <10 kbp overlap
as second round for sequencing
Screen deeper library if no BACs
overlap by <10 kbp on an end
Screen alternate libraries if no BACs
overlap by < 30 kbp

Sequencing by Project 6/96-2/97



Summary

Category	# of BACs	Size
Submitted to GenBank	18	2,643,073
Closure	5	735,000
Random	2	360,000
Ready for random	12	1,875,000

Library Team

Cheryl Phillips, Kun Shen, Marie LaBombard

Random Team

10 377xl, 9 373, 1 373xl, 5 Catalyst

Joyce Fuhrmann, Tanya Mason, Steve Bass, Paul Sadow, Jen Tench, Lisa Jiang, Roy Sittig

Closure Team

Rhonda Brandon, Kurt d'Andrea, Sean Sykes, Tracy Spriggs, Tammy Lockwood

Gene List

G1 to S Phase transition protein 1, GST1
B cell maturation protein
hypothetical protein CIT987SK_2A8_1
extoses like gene (partial)
hypothetical protein CIT987SK_362G6_1
hypothetical protein CIT987SK_362G6_2
T-complex protein 1, Beta subunit (TCP-1-BETA), partial
Human gene for Myosin heavy chain (partial)
Multidrug resistance-associated protein isolog
Multidrug resistance-associated protein
pM5
eIF-3 p110 subunit

12 genes 2,363,073 bp

OR

1 gene per 196 kbp (!)

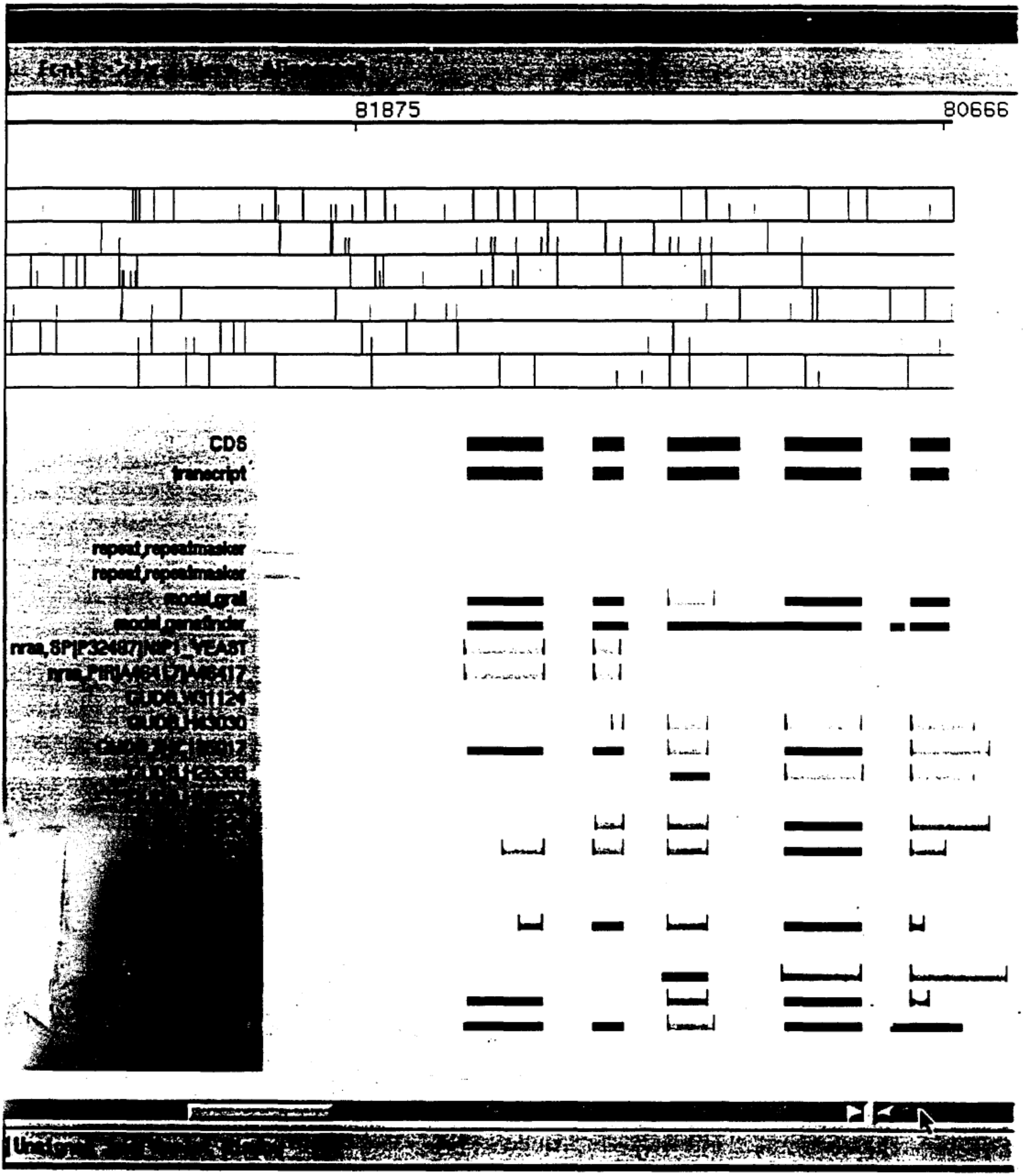
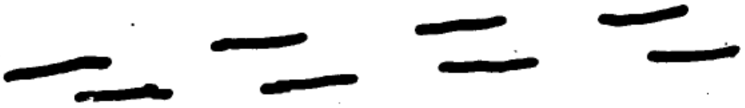

81875
80666
CDS
transcript
repeat,repeatmasker
repeat,repeatmasker
model,grail
model,genefinder
nraa,SP|P32487|NP1_YEAST
nraa,PIR|A48417|A48417
GUDB,W31124
GUDB,H63130
GUDB,H23368

Table of Double Chemistry Effort and Results

BAC name	Total Len	Single-strand	Terminators	Terms in area	Bases Changed
C16Q	227,403	25,110	209	74	3
CPBA	136,182	14,991	202	73	2

DAVID COX

Stanford Human Genome Center



Chromo 21

EPM1 1.2 mb

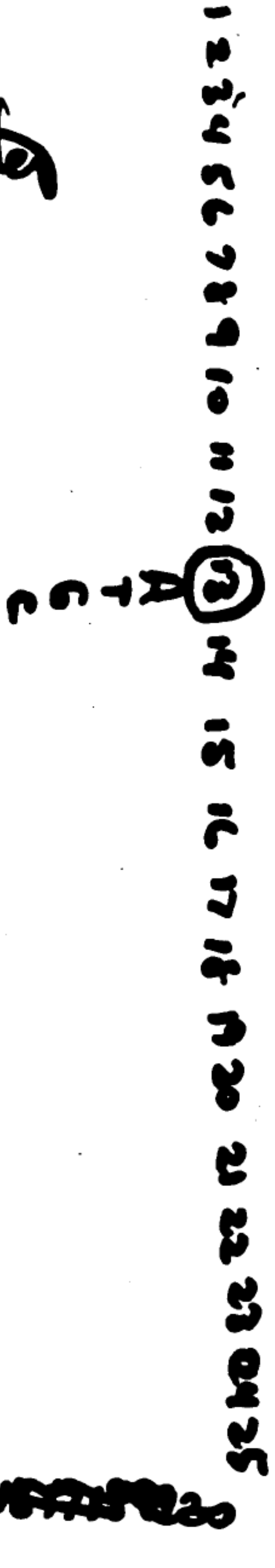
DS .4 mb

Chromo 4

4q25 5 mb

Finished 100kb

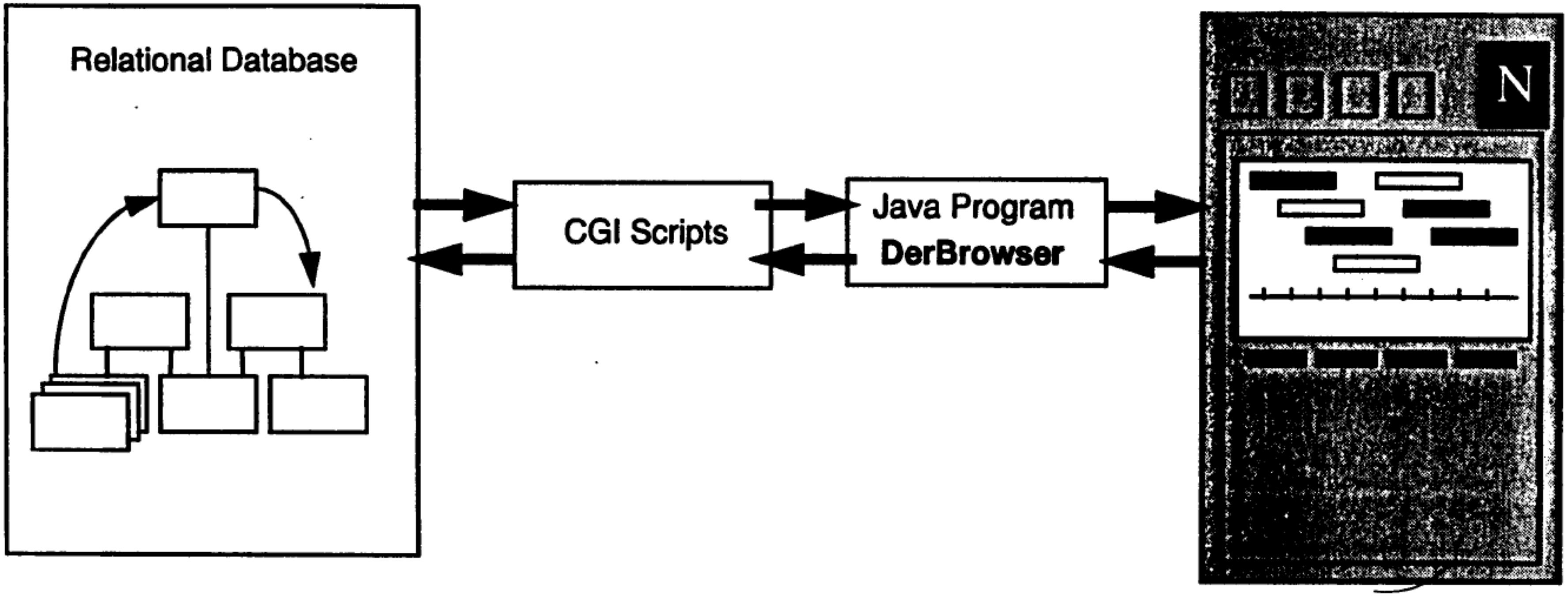
In GenBank 1.2mb >3kb



~~1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20~~

FIONA FRANCIS
(LEHRACH LAB)

Web access to sequencing status



Relational Database

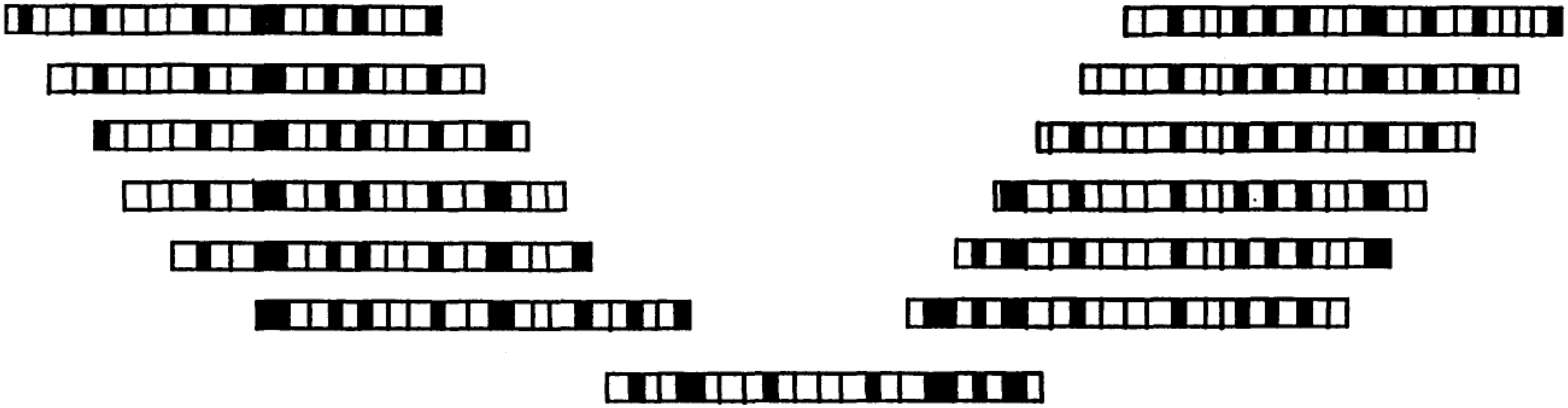
CGI Scripts

Java Program
DerBrowser

DerBrowser:

http://www.mpimg-berlin-dahlem.mpg.de/~andy

Preselection of shotgun clones



Projects completed

Xp22

+ Region: DXS8254 - DXS1683, containing the PEX gene
+ Size: 243 kb, contiguous
+ Status: complete
+ Accession number: Y10196

Projects in progress

21q22.3

+ Region: D21S349 - MX1
+ Size: 500 kb
+ Status: 3 cosmids and 2 PACs at different sequencing stages, other shotgun libraries in preparation

Xq28

+ Region: DXS304 - DXS1345, proximal to MTM gene
+ Size: 320 kb (one cosmid in region sequenced previously)
+ Status: finishing stage

Xq13

+ Region: GJB1 - DXS559
+ Size: 500 kb
+ Status: shotgun libraries in preparation

Xq13

+ Region: DXS227
+ Size: 150 kb
+ Status: shotgun libraries in preparation

Xq12

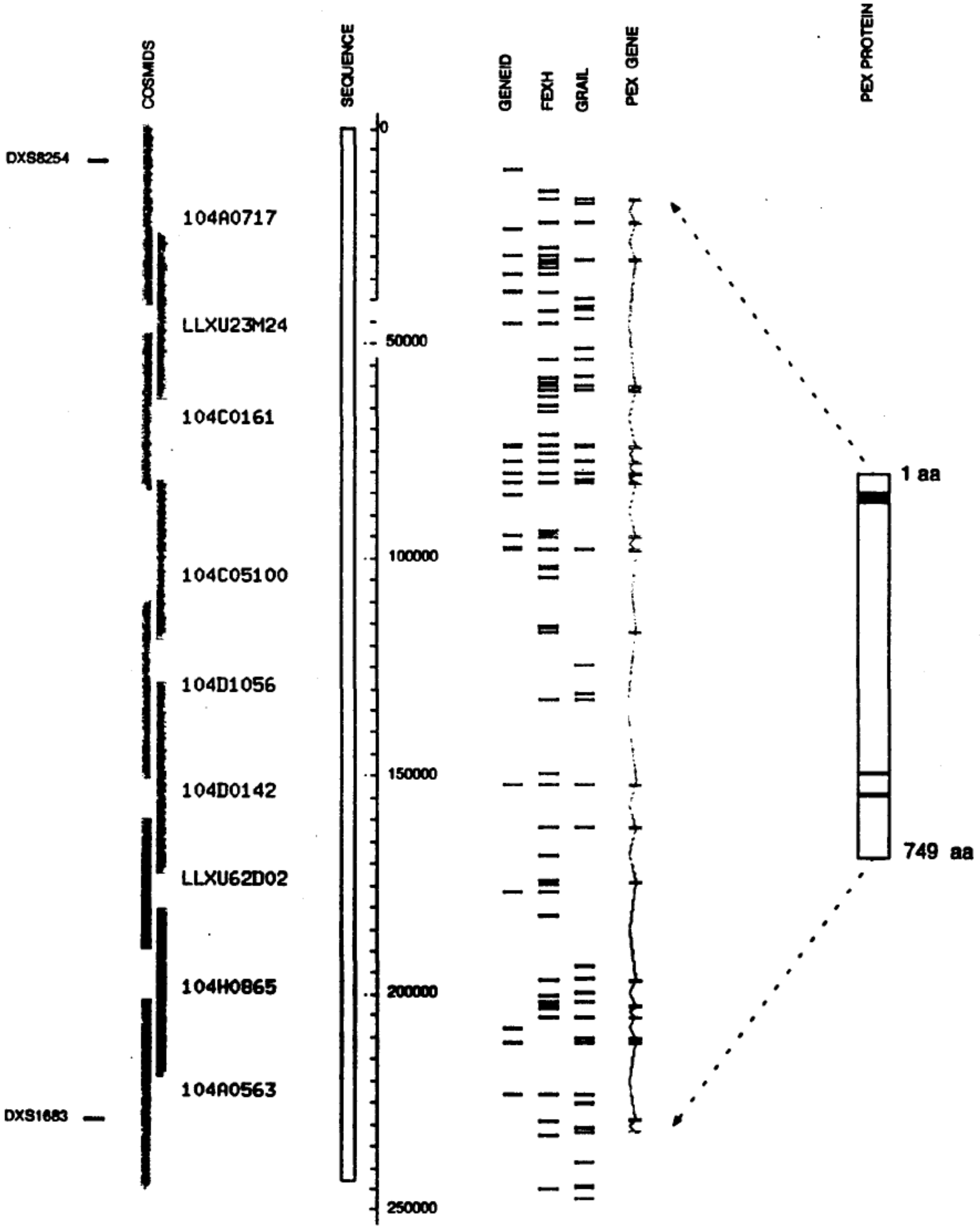
+ Region: DXS908
+ Size: 150 kb
+ Status: shotgun libraries in preparation

17p11

+ Region: D17S71 - D17S58
+ Size: 1000 kb
+ Status: shotgun libraries in preparation

Data quality

- Attempt to close all gaps
- Double stranding/alternative chemistry
- Cover all regions by sequence from more than one shotgun clone
- Attempt to resolve all problematic regions
- Confirm sequence by comparison to restriction digests



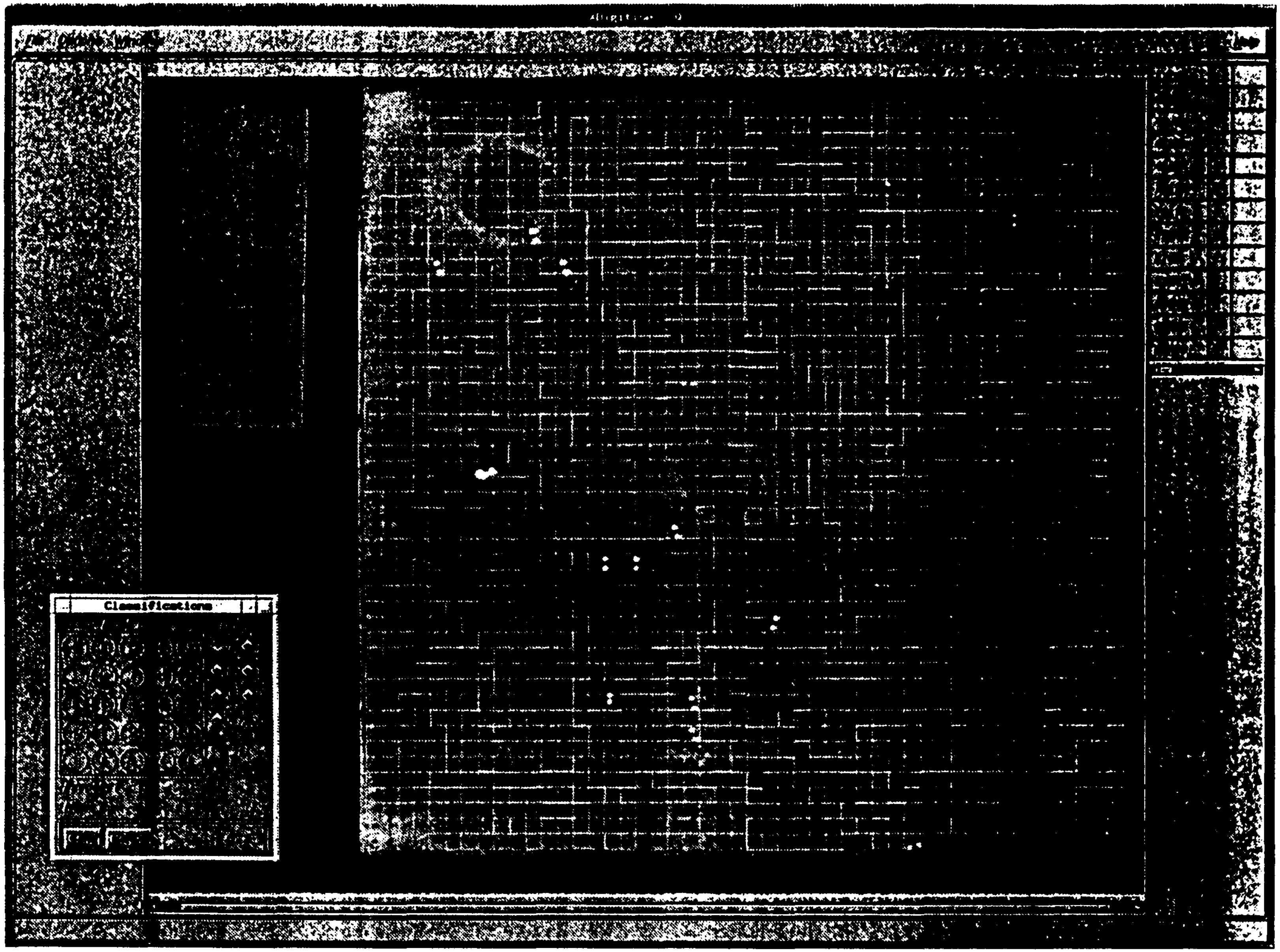
Accession number: Y10196

Assembly and analysis of sequence

- Staden package: pregap programs, xgap and gap4
- Phred/Phrap (P. Green) and Phrap2Gap (Sanger Centre)
- Gene prediction: Grail, Genefinder (V. Solovyev), Xpound
- Masking of repeats: Repeat Masker/Repbase (A. Schmidt) and Blastn/Simple.db with XBLAST (J-M Claverie)
- Database searches: Blastn and Blastx/ nr and dbEST
- Search and analysis tools: Seqsplit/Blastunsplit and MSPcrunch/Blixem (E. Sonnhammer and R. Durbin)
- Data storage and visualisation: Acedb

Shotgun cloning and sequencing

- Starting DNA: CsCl purified cosmid/PAC
- Standard shotgun cloning: insert sizes 1.2-1.8 kb, sequencing vector: pUC18
- Clones picked in microtitre dishes, inserts PCR amplified
- Cycle sequencing performed using ABI Catalyst, reactions run on ABI 377s
- Data collection, transfer to Unix environment
- Gap closure/finishing after assembly: reverse reads, directed primer walking, PCR


Classifications

Chr. 21 - construction of sequence-ready maps

- Libraries: Chr. 21 cosmid and whole genome PAC (and BAC)
- Hybridisation screening using STS probes and riboprobes (extension of existing contigs, and anchoring of new ones)
- FISH mapping of selected clones
- Contigs also contributed by collaborating groups
- Restriction digests to aid selection of a minimal tiling path
- Higher resolution fingerprinting performed in selected regions
- End sequencing of clones to aid gap closure

JEAN
WEISSENBACH

French Sequencing Center

Centre National de Séquençage

Budget	14 M $
Staff	110-120
Location	Evry (near Généthon)
Starting	Summer 1997
Projects	To be submitted

Project Evaluation

Scientific Committee

- scientific quality
- feasibility, opportunity
- scientific interest
- scientific priority

Steering Committee

- political recommendations about projects
- priority decisions
- recommendations about policies on data release and intellectual property.

JOHN
MATTICK

THE AUSTRALIAN GENOME RESEARCH FACILITY (AGRF)

- funded at $A10m ($US8m) for equipment only (project funding to be obtained separately)*
- TWO DIVISIONS:
  (1) DNA SEQUENCING at the Centre for Molecular and Cellular Biology, University of Queensland, Brisbane
  (2) DNA GENOTYPING at the Walter and Eliza Hall Institute of Medical Research, Melbourne
- currently in final stages of planning and equipment acquisition, due to begin operations mid-1997 (~ 30 × 377s + assoc. equipment, robotics)

DNA SEQUENCING (University of Queensland)

Current status: 4 × 373s ~ 800 templates/week

Projected: ~ 15 × 377s (+ existing 373s)
~ 1500 - 2,000 templates/day

Housing: Proposed new Institute $A50m ($US40m)
- have obtained $A30m from University of Queensland and State Government
- attempting to raise $A20m from Federal Government and other sources
- construction 1997-1999 with temporary housing for facility in the interim

.../2

OPERATIONAL

- AGRF will be a generic high-throughput sequencing facility, not restricted to particular projects, available to Australian and regional research community.
- no operational funds (yet) voted to the facility. These are intended to be derived by participating groups from granting agencies.
- two modes of operation:
  - (a) contract/service — on behalf of client groups who will supply funds and who will take primary responsibility for cloning, library construction, sequence assembly and annotation (using own facilities, supported by services provided by AGRF and ANGIS – Australian National Genome Information Service)
  - (b) bid for specific funds to undertake large projects in-house, and construction of teams for this

FUNDING*

- from existing granting agencies
- working to convince Australian Government to set up a specific fund (~ $A20m/year) to support genome-scale projects, including an Australian participation in the human genome sequencing project.

ANDRE ROSENTHAL

technology

1997   20 ABI 377
- 4 (IMB)
- 8 (BMBF-BEO)
- 8 (BMBF-DLR)

(96' 1.300 reads/day
97' 3.000 reads/day)

6 production groups (1 Postdoc, 3 technicans)

1 bioinformatics group (5 people)

1 library group (1 Postdoc, 3 technicans)

production (6 groups: 1 Postdoc, 3 TA's)

- picking, preping, sequencing, loading, data transfer, assembly, finishing, annotation

1997	1998	1999
6 x 1 Mb = 6 Mb	6 x 2 Mb = 12 Mb	6 x 3 Mb = 18

funding

- Land Thuringia (renting lab space, lab furniture)
- federal government BMBF-BEO 13 Mill D
  1995-2000
- federal government BMBF-DLR 14 Mill
  (May 97 - April 2.000) (30 cents/ba

Resources

cosmids, PAC's, BAC's

targets	(1997-2000)		maps
X	Xq 28	3 Mb	- Nelson/Gibbs - Poustka Kioschis (Heidelberg
	Xp 11.23 Xp 11. 4	2.5 Mb	- Meindl (Munich)
	PAP1	1 Mb	- Rappold (Heidelberg)
21q		28 Mb	- Yaspo (Berlin) - internat. chr. 21 consort
7	7q 22	7 Mb	- Scherer Tsui (Toronto)
	7q 32	0.5 Mb	
mouse	syntenic to Xq 28	3 Mb	

Genome Sequencing Centre at IMB, Jena (Germa

1996 2.5 Mb completed ↗ 1.5 Mb Genbanl

↘ 1 Mb annotati
phase

1997 - 2000 (April)

Σ 40 Mb

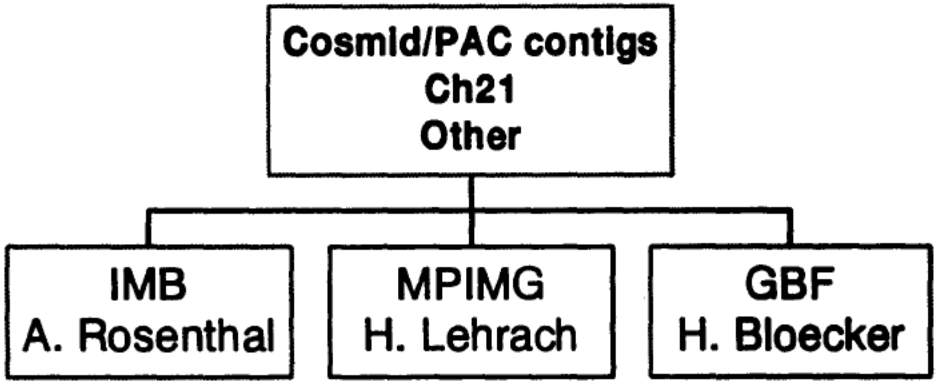
↙ 28 Mb (BMBF-DLR State Thuringia)

↘ 12 Mb (BMBF-BEO State Thuringia)

1997	6 Mb	(4 + 2)
1998	12 Mb	(9 + 3)
1999-	19 Mb	(15 + 4)
2000 (Jan-April)	3 Mb	

German Human Genome Project

Genomic sequence analysis of human chromosome 21 and selected regions of the human genome



Cosmid/PAC contigs
Ch21
Other

IMB
A. Rosenthal

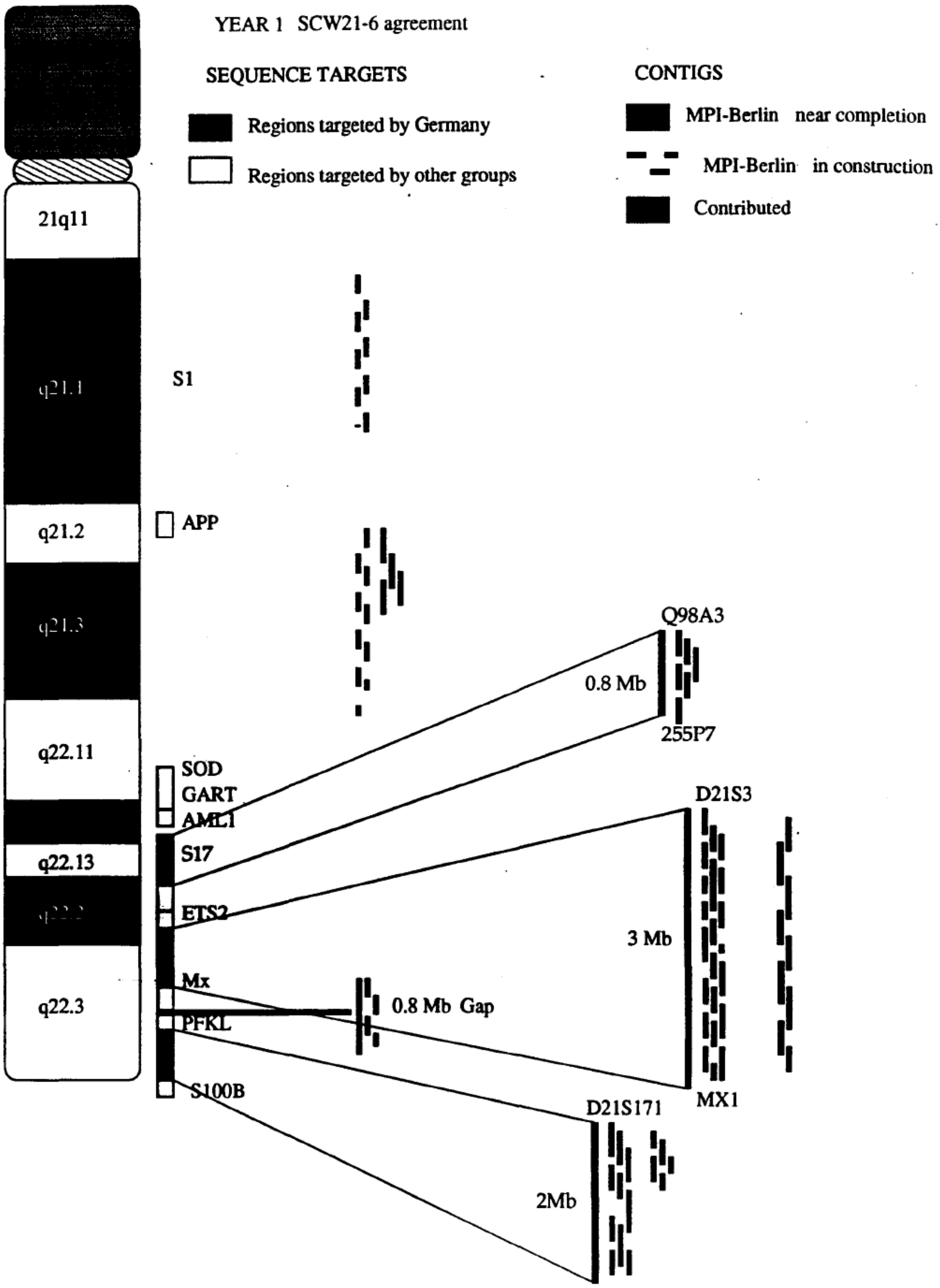
MPIMG
H. Lehrach

GBF
H. Bloecker

Coordinator

year 1	4Mb	1Mb	1Mb
year 2	9Mb	2Mb	2Mb
year 3	15Mb	3Mb	3Mb

YEAR 1 SCW21-6 agreement



PHIL GREEN

GENOME SEQUENCE QUALITY CRITERIA

- Fidelity:
  - *2X validation* of all sequence-ready clones, using method adequate to detect small ($< 1kb$) coligations, deletions, transposon insertions
- Accuracy:
  - Error rate $< 1/10kb$
  - Base-specific error probabilities submitted with sequence
  - Independent test of assembly accuracy
- Contiguity:
  - All gap sizes estimated
  - All sequence contigs *oriented* and *ordered* within the chromosome

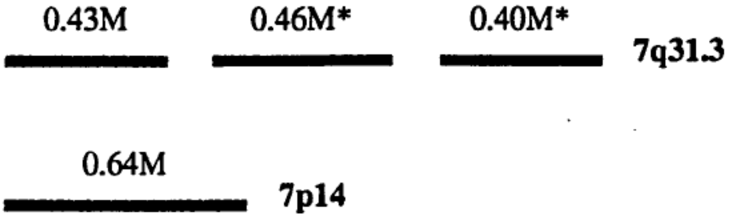
UWGC SEQUENCING STRATEGY: KEY FEATURES

- MCD mapping
  - Clone validation
  - Better tiling paths
  - More efficient finishing (gap closure)
  - Assembly verification
  - Current cost: $.05 to $.12 per bp
- Long reads
  - Reduce finishing and assembly problems
  - Raise machine costs, lower all other costs
- Objective finishing criteria based on error probabilities

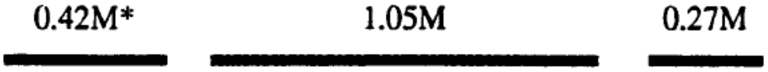
UNIV. OF WASHINGTON GENOME CENTER MAMMALIAN SEQUENCING PROJECTS

- Human chromosome 7
- Human HLA Class I
- Mouse T-cell receptor alpha

Human chromosome 7



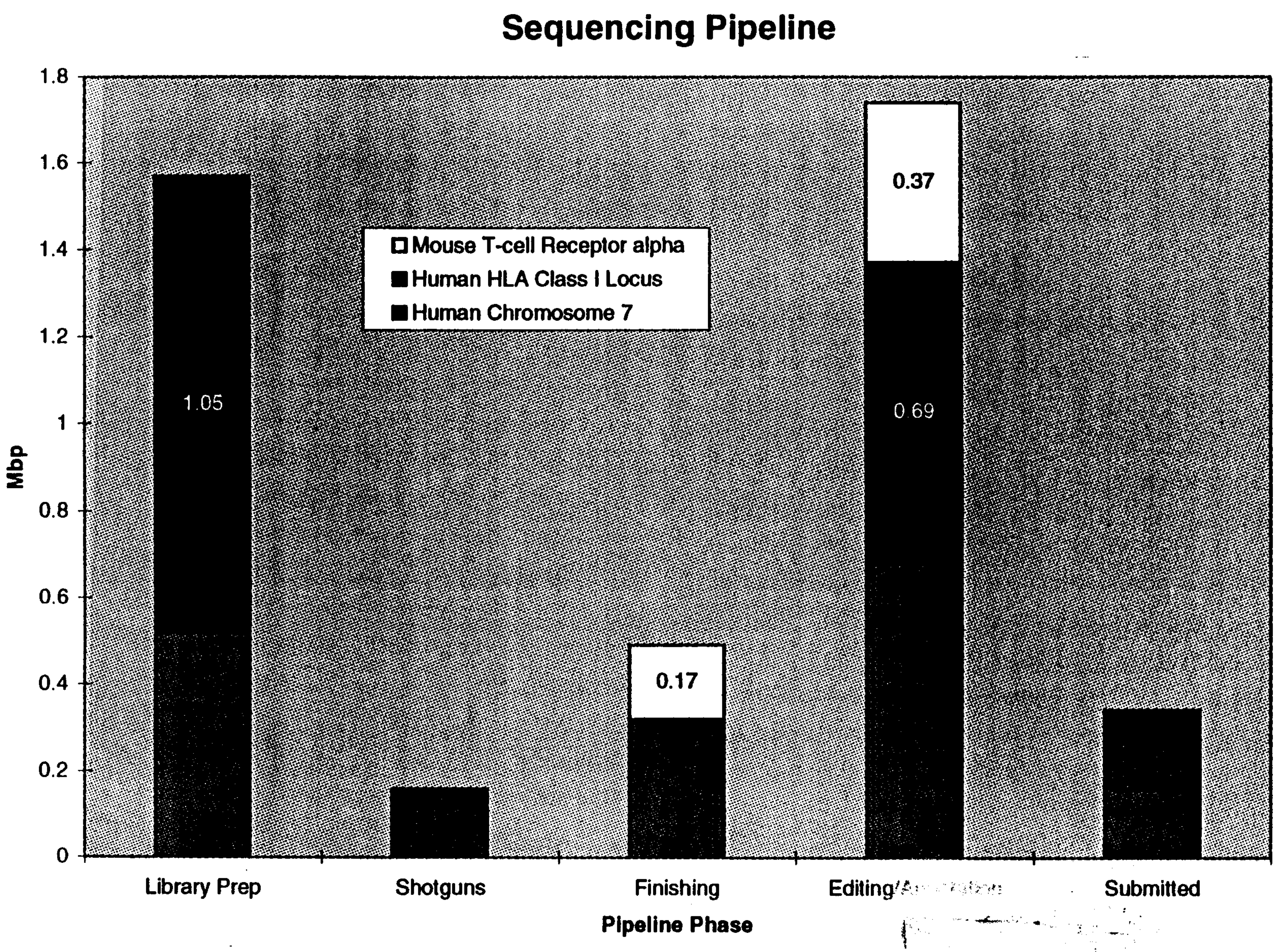
Human HLA class-I locus



Mouse T-cell receptor α

This 1M region is covered by 75 BACs and is being MCD mapped by a combination of BAC-to-cosmid subcloning and direct BAC fingerprints. The details are in another figure.

Color code: ■ mostly sequenced, ■ being sequenced


Sequencing Pipeline
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
Mbp
□ Mouse T-cell Receptor alpha
■ Human HLA Class I Locus
■ Human Chromosome 7
1.05
0.17
0.37
0.69
Library Prep
Shotguns
Finishing
Editing/Annotation
Submitted
Pipeline Phase

INTERNAL ACCURACY ASSESSMENT

- **MCD mapping.**

  Test: MCD maps are compared to sequence-predicted maps.

  Results:

  - **No mapping errors thus far in HLA and Chr. 7 regions (1.2*Mb* finished sequence).**

- **Sequencing.**

  Test: All cosmids are independently finished, and sequences of overlapping same-haplotype cosmids are compared.

  Results:

  - **Chr. 7:**
    0 discrepancies in 2 X 38802 bp
  - **HLA:**
    2 discrepancies in 2 X 43084 bp
    - **1 mismatch (phrap error – incorrect read selected)**
    - **1 apparent cosmid mutation (12 bp insertion/deletio in repeat region)**

MCD MAPPING

- Start with large clones (YACs or BACs) from region of interest; 2X depth
- Subclone into cosmids; 20-30X depth
- Restriction digests with three enzymes
- Construct map of restriction sites & clone ends

sWSS106
Segmap V. 3.40 Data File Date: Wed Oct 30 10:54:22 1996
Chromosome 7 q31.1 q34 Uncomputed Map 100 kb/cm

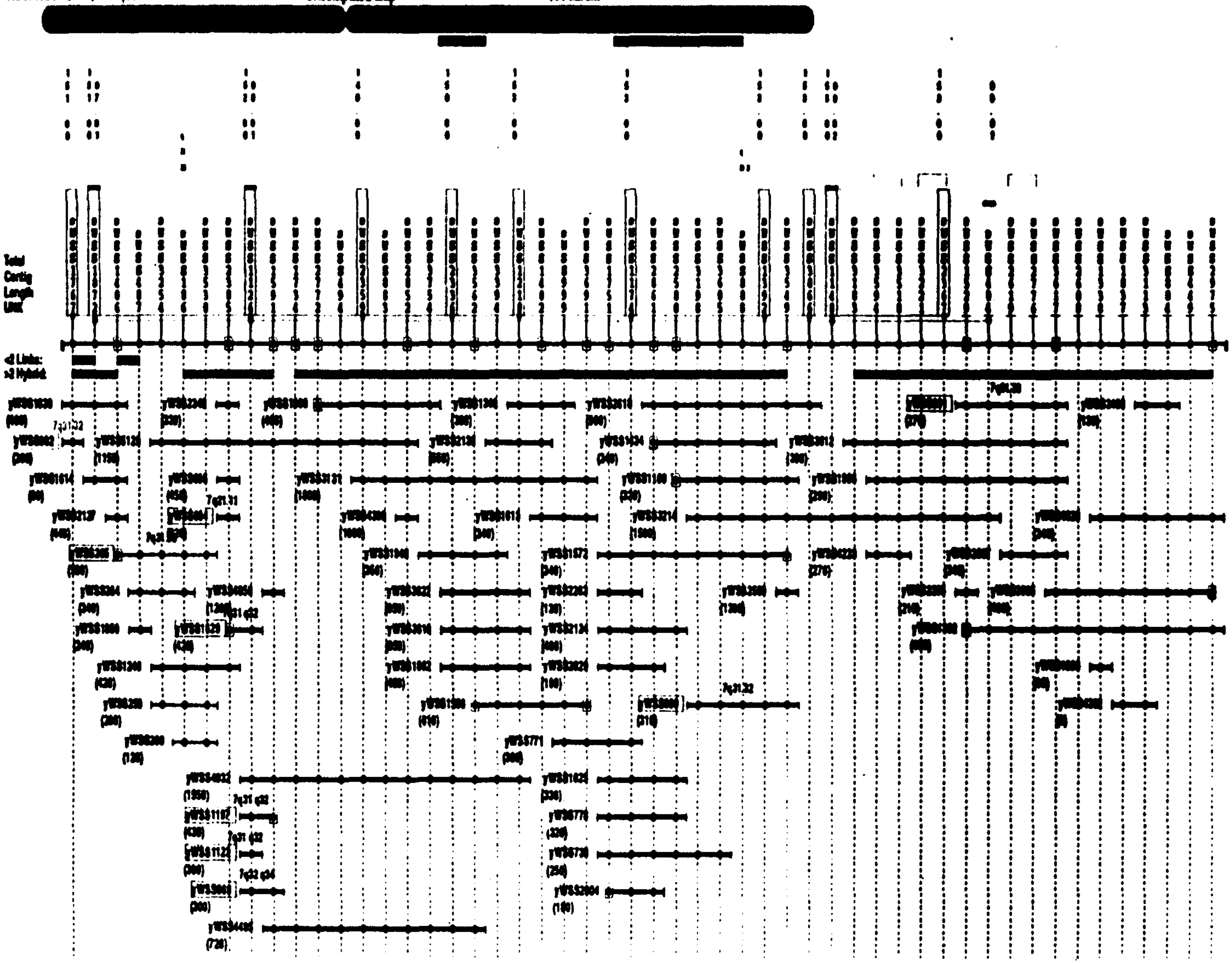


Table 1. Summary of YAC→cosmid MCD maps for portions of human chromosome 7.

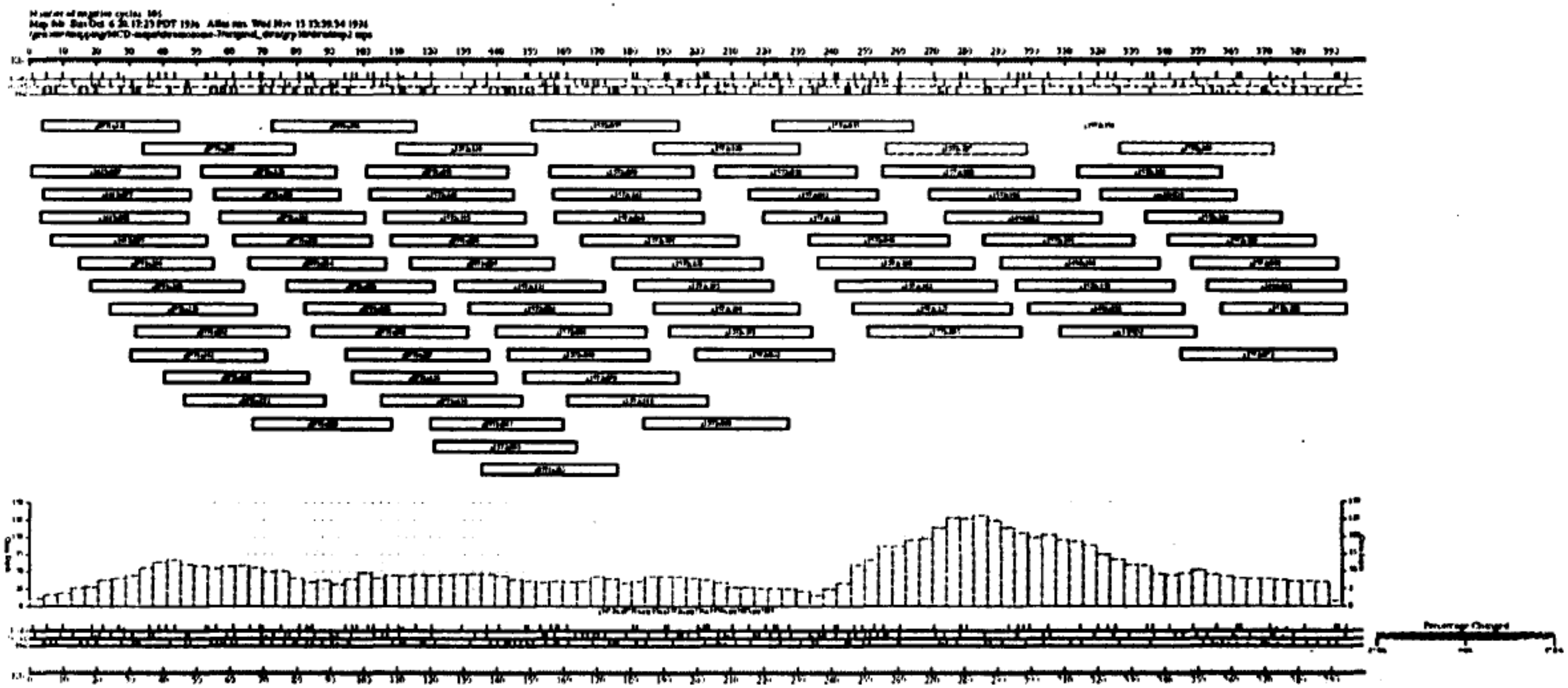
Chr-7 YACs	Coverage [a]	N_f [b]	N_f [b]	N_f [b]	Co-ligations [c]	Map Size [d]
		(EcoRI)	(HindIII)	(NsiI)		(Kbp)
yWSS771	30.3	9.8 / 1.2	8.4 / 1.2	11.4 / 1.2	2.8%	44+170
yWSS1346	29.2	10.5 / 1.2	12.4 / 1.3	10.0 / 1.3	3.0%	281
yWSS1434	20.5	7.4 / 1.3	6.8 / 1.4	7.4 / 1.6	7.8%	156
yWSS1564	16.7	9.2 / 1.3	10.4 / 1.5	9.8 / 1.3	7.9%	640
yWSS1572	31.5	8.0 / 1.2	9.1 / 1.2	9.0 / 1.3	4.5%	292
yWSS1613	26.3	10.6 / 1.2	10.6 / 1.1	11.5 / 1.3	3.5%	136+56
yWSS1862	23.4	8.4 / 1.2	11.0 / 1.2	11.6 / 1.3	3.4%	261
yWSS1980	20.7	8.3 / 1.1	8.5 / 1.1	10.8 / 1.1	5.7%	278

[a] Coverage is calculated assuming a 40 Kbp insert size. Clones left out of the map because they could not be uniquely placed are included in the coverage calculation, while co-ligations and yeast impurities are not. When there are two contigs, we simply add their sizes to compute the coverage.

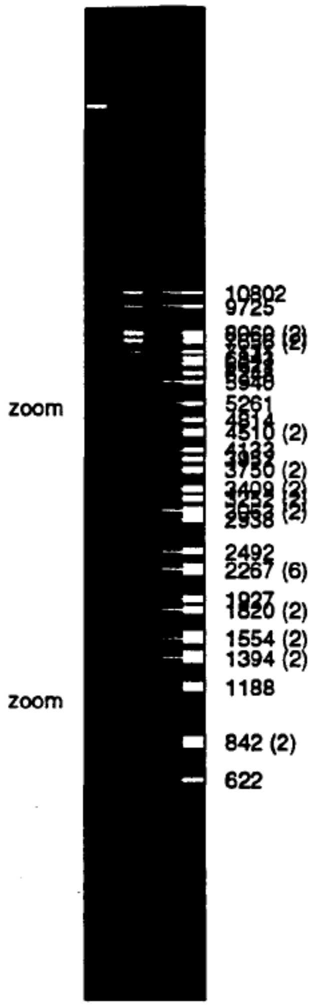
[b] N_f refers to the average number of fragments observed in a clone, which is the first number given in each row. The second number indicates the average number of fragments per fragment group, an indication of how well ordered the restriction fragments are in the maps. Contigs smaller than 100 Kbp are not included when summarizing fragments per fragment group.

[c] Co-ligations are cosmids that contain both a human insert from the targeted region and an unrelated piece of DNA that is inserted between the end of the human insert and the cosmid vector.

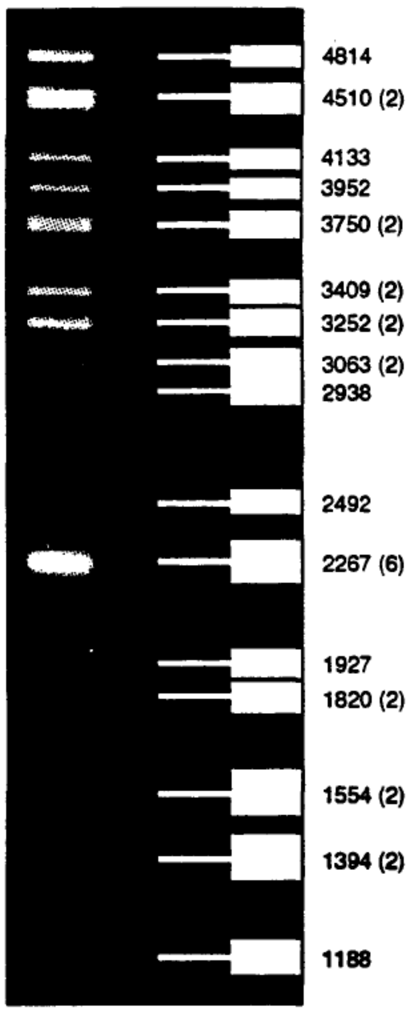
[d] Map sizes are based on the sum of the restriction-fragment sizes. The gap in the overlap region between YACs yWSS771 and yWSS1613 has not yet been closed. These maps agree perfectly with each other on either side of the gap, and both maps stop at exactly the same places.



■ Library Prep ■ Shotguns □ Finishing ▨ Editing/Annotation ■ Submitted lane number 28 on file "ma13"



ma24 (HindIII)



zoom subregion

```
    -------- EcoRI --------     ------- HindIII -------     -------- NsiI ---------
       MCD map         Clone       MCD map         Clone       MCD map         Clone

         :                           :                           :
    ----------                      691.28                  ----------
       2084.47                  ----------                     4799.00
    ----------                     4268.57                  ----------
       1122.77                  ----------                     1561.94
    ----------                     1104.83                     2559.12
       5079.10                  ----------                  ----------
    ----------                     1800.14                     9148.59
       1123.18                  ----------                  ----------    ----------
    ----------                     1973.64                     5048.52       5052.00
       1273.74                  ----------                       ???       * 5709.00
    ----------                     1858.81                  ----------
       9915.76                  ----------                     1575.81       1586.00
    ----------    ----------       3876.31                  ----------
       3465.80       3462.00    ----------                     3378.90       3378.00
         ???       *13976.00        974.84                  ----------
    ----------                  ----------                   * 4335.94         ???
       1673.51       1676.00       2944.00                  ----------
    ----------                  ----------                     6350.05       6343.00
       3330.49       3327.00       5435.75                  ----------
    ----------                  ----------    ----------       2141.53       2146.00
       1221.67       1223.00       4864.69       4860.00    ----------
    ----------                       ???       * 8374.00       6769.09       6762.00
     *12709.65         ???      ----------                  ----------
    ----------                     1550.16       1550.00        630.39        629.00
       9836.67       9778.00    ----------                  ----------    ----------
    ----------                      768.44        768.00      10373.77
       1049.14       1052.00    ----------                  ----------
    ----------    ----------       1111.40       1111.00       1582.33
       4244.29                  ----------                  ----------
    ----------                   * 6975.33         ???        14942.58
       3008.12                  ----------                  ----------
    ----------                     2127.50       2130.00       1222.25
       7014.18                  ----------                  ----------
    ----------                     2769.91       2770.00        970.00
       3112.29                  ----------                     4153.25
    ----------                     1789.55       1791.00    ----------
       5941.43                     1355.84       1353.00       2833.33
    ----------                  ----------                  ----------
       2019.67                     1553.21       1550.00       6344.00
       8330.00                  ----------                  ----------
    ----------                     2300.24       2301.00        961.00
       2650.00                  ----------                     3832.00
    ----------                     7324.61       7304.00    ----------
       3514.00                  ----------    ----------       1364.67
    ----------                     7077.58                     1755.33
       2361.40                  ----------                     4019.83
    ----------                     8837.62                     5315.67
        842.83                  ----------                  ----------
       1113.00                     1695.92                     7826.00
       4335.00                     3706.42                      797.40
    ----------                  ----------                     1693.60
         :                           :                           :
```

A transposon-insertion detected on chromosome-7 yWSS1346. Every enzyme domain in the aberrant clone has one extraneous fragment that cannot be matched to the MCD map. However, if something like 1400-bp is subtracted from each of these 3 extraneous fragments, the clone can be mapped in.

OBJECTIVE PROCEDURE TO ACHIEVE DEFINED ERROR RATE

- Following shotgun assembly, estimate error probability at each consensus base position; compute expected number of errors for entire cosmid or BAC.
- Finishing: collect enough additional data, or edit, in regions of highest error probability ("gaps") to force expected number of errors below 1 per 10 kb.
- Periodically, for selected cosmids, test agreement between expected number of errors and actual number of errors (relative to "gold standard").

- Monitor raw data quality using per read distribution of error probabilities.
- Explore optimal (least expensive) shotgun / finishing tradeoff yielding target error rate.

CURRENT TECHNOLOGY DEVELOPMENT

- MCD mapping:
  - BAC restriction digests
  - Automated clone anomaly detection
- Sequence assembly and editing:
  - Phrap: Improved error probabilities, resolution of large exact repeats, use of map information, reassembly directives
  - Phred: Lane processing, compression resolution
  - Consed: Tags, custom navigation

UNIVERSITY OF WASHINGTON GENOME CENTER
Maynard V. Olson, Director

MCD Mapping

Jun Yu, Leader	(ft)
Ying Ge	(ft)
Zahra Magness	(ft)
Ruolan Qiu	(ft)
Channakhone Saenphimmachak	(ft)

Sequencing

Shawn Iadonato, Leader	(ft)
Cindy Desmarais	(ft)
Thomas Gilbert	(ft)
Kim Harris	(ft)
Lloyd Lytle	(ft)
Oanh Nguyen	(pt)
Quynh Pham	(ft)
Karen Phelps	(pt)
Steven Swartzell	(ft)

Software Development

Brent Ewing	(ft)
David Gordon	(ft)
Arian Smit	(ft)
Ed Thayer	(ft)
Colin Wilson	(ft)

Production Informatics/Map Finishing

Gane Wong (ft) and Charles Magness (pt)

Kerry Bubb	(ft)
Jina Chang	(pt)

UNIVERSITY OF WASHINGTON GENOME CENTER

Maynard V. Olson, Director

Collaborators:

NHGRI:	Eric Green
Fred Hutch Cancer Research Center:	Dan Geraghty, Thierry Guillaudeux, Marta Janer
University of Washington - Molecular Biotechnology:	Leroy Hood, Inyoul Lee, Lee Rowen
University of Washington - Computer Science:	Richard Karp
Washington University - Computer Science:	Will Gillett, Liz Hanks

ELLSON citation

PRODUCTION SEQUENCING OF MAMMALIAN DNA BY ORDERED SHOTGUN SEQUENCING (OSS) STRATEGY

[1]Peter Ma, [1]Chun-Nan Chen, [1]Ying Su, [1]Primo Baybayan, [1]Aleli Siruno, [1]Jeanette Evans, [2]Richard Mazzarella, [2]David Schlessinger and [1]Ellson Chen
[1]Advanced Center for Genetic Technology, Applied Biosystems Division of Perkin Elmer Corp., 850 Lincoln Center Drive, Foster City, CA 94404, and [2]Department of Molecular Microbiology, Washington University School of Medicine, St. Louis MO 63110.

Ordered shotgun sequencing (OSS) has been successfully carried out to sequence over 2.3 megabases DNA (>20 large-insert clones) from human X-chromosome isochores with different GC levels. The approach combines mapping and sequencing of YACs, BACs, or PACs with a hierarchical strategy that incorporates a feedback loop [Chen, E. et al., Genomics 17, 651-656 (1993); Chen,C et al., Nucleic Acids Res, 24, 4034-4041 (1996)]. Clones are recovered by STS-based screening of clones (see Williams et al., these ABSTRACTS). The method starts by randomly fragmenting a BAC, YAC or PAC to 8-12 kb pieces and subcloning those into lambda phage. Insert-ends of these clones are sequenced and overlapped to create a partial map. Complete sequencing is then done on a minimal tiling path of selected subclones.

OSS is currently delivering sequence at a cost comparable to methods that have been established far longer. Automation is facilitated by adapting PCR to prepare all sequencing templates, along with further improvements in sequencing technology and informatics. The approach also provides considerable flexibility in the choice of sequencing substrates. For example, subclones containing contaminating DNA can be recognized and ignored with minimal sequencing effort; regions overlapping a neighboring clone already sequenced need not be redone; and segments containing tandem repeats or long repetitive sequences can be spotted early on for targeted handling.

The encouraging results have led to an expanded goal of increasingly cost-effective genomic sequencing of 35 megabases, initiated on portions of Xq26 (1.5 Mb), Xq27 (1.5 Mb), Xp11.2 (1 Mb), Xq 12 - q21 (17.5 Mb), Xq21.3 (4.5 Mb); chromosome 3 (10 Mb, primariily in 3p21); and comparative sequencing of 8 Mb of mouse DNA, including the t-complex In1 and In2 regions (and corresponding human 6q24-q27), and segments homologous to Xp11.2 and Xq13 DNA already in process.

PROSPECTS

Feb, 97

(Production sequencing at PE-ABD/WU Genome Center)

Short-term (in 1997), 3Mb finished sequences (in addition to 2 Mb finished as of 12/31/96) on portions of:

- 1 Mb in Xp11.2, from DXS1008E to DXS423E.
- 1 Mb in Xq13.2, from DXS227 to DXS7025E
- 1 Mb in Xq26, from GPC3 to DXS8033.
- 1.5 Mb in Xq27, from F9 to DXS984.

Long-term (by 2000), >30 Mb sequences on:

- 13 Mb region in Xq11.2-q13.2 from DXS1 to DXS441 (about 2 Mb of which is being sequenced so far),
- 4.5 Mb Xq21.3 XY homology region, from DXS1217/DXYS1X to DXS3.
- 10 Mb of chromosome 3p21 and selected BACs from 3q23 and 3q29.
- 8 Mb of mouse DNA, including the xce locus and inversion regions In1 and In2 of the t-complex (as well as the corresponding human 6q24-q27).

Status of Hun [illegible] mosome Sequencing at ACGT

	Locus	"MB" Index	Clone type	Insert size (Kb)	ABD Project #	Sequence Region (Marker Limits)	Status	Kb done	Remarks	STS Content: sWXDs
1	q28	158	9 cosmids	220	*	CV-G6PD	completed	220		
2	q24-25	118	yWXD703	135	A&B	ANT2	completed	135		
3	q13	67	bWXD161	220	C	DXS227 -7025E	1 gap	210		
4	q26	131	bWXD8	165	D	GPC3-DXS8033	completed	165		2271, 1455
5	q13	67	bWXD3	95	E	DXS227 -7025E	completed	95	submit	
6	q26	131-132	bWXD9	250	F	GPC3-DXS8033	1 gap	240		
7	q25	124-136	pWXD6	110	G	DXS100	1 gap	132		
8	q25	124-136	pWXD1	100	H	DXS7831	3 gaps	95	too many gaps	
9	q13.2	72-74	bWXD27	135	J	DXS227 to 7025E	1 gap	135		905,3679,3678,15, 981, 599
10	q13.2	72-74	bWXD40	103	K	DXS227 to 7025E	final checking	138		515
11	q13.2	72-74	bWXD42	110	L	DXS227 to 7025E	final checking	99		1254, 1253, 870
12	q13.2	72-74	bWXD14	112	M	DXS8066 - DXS1221	2 contigs	81		1255, 2891
13	q13.2	72-74	bWXD20	102	N*	DXS1679 ?	14 contigs	100		2870
14	q13.2	72-74	bWXD36	177	O	DXS8066 - DXS1221	2 contigs	151		1875
15	p11.2	54-56	bWXD142	140	P	OATL2-CEN	2 contigs	89		1995, 3675, 2559
16	p11.2	54-56	bWXD111	109	Q	OATL2-CEN	2 contigs	70		570, 3676, 2106, 3527, 3525, 1977, 2183, 2894, 2107, 1118
17	p11.2	54-56	bWXD137	158	R	OATL2-CEN	starting			2560, 1977
18	q27	139-140	bWXD90	77	S	DXS1192-DXS119	2 contigs	57		1445
19	q27	139-140	bWXD100	152	T	DXS1192-DXS119	2 contigs	64	Tough to PCR	2623
20	q27	139-140	bWXD105	124	U	DXS1192-DXS119	1 contig	65	Tough to PCR	2462
21	q26	131-132	bWXD168	121	V	GPC3-DXS8033	on hold			415, 27
22	q26	131-132	bWXD171	160	W	GPC3-DXS8033	on hold		33% coli?	791, 2457
23	q26	131-132	bWXD173	179	X	GPC3-DXS8033				791, 1319
24	q26	131-132	bWXD180	160	Y	GPC3-DXS8033				1307, 1334, 415
25	q26	131-132	bWXD181	160	Z	GPC3-DXS8033	Cancel?		overlap with D	1151, 1455, 2271, 2863
26	q26	131-132	bWXD200	240	AA	GPC3-DXS8033				1182, 1928, 2457
27	q26	131-132	bWXD177	92	AB*	GPC3-DXS8033	by shotgun			1151, 385
			total	3906				2341		

Total finished 2,341 Kb on 2/12/97

* by shotgun sequencing

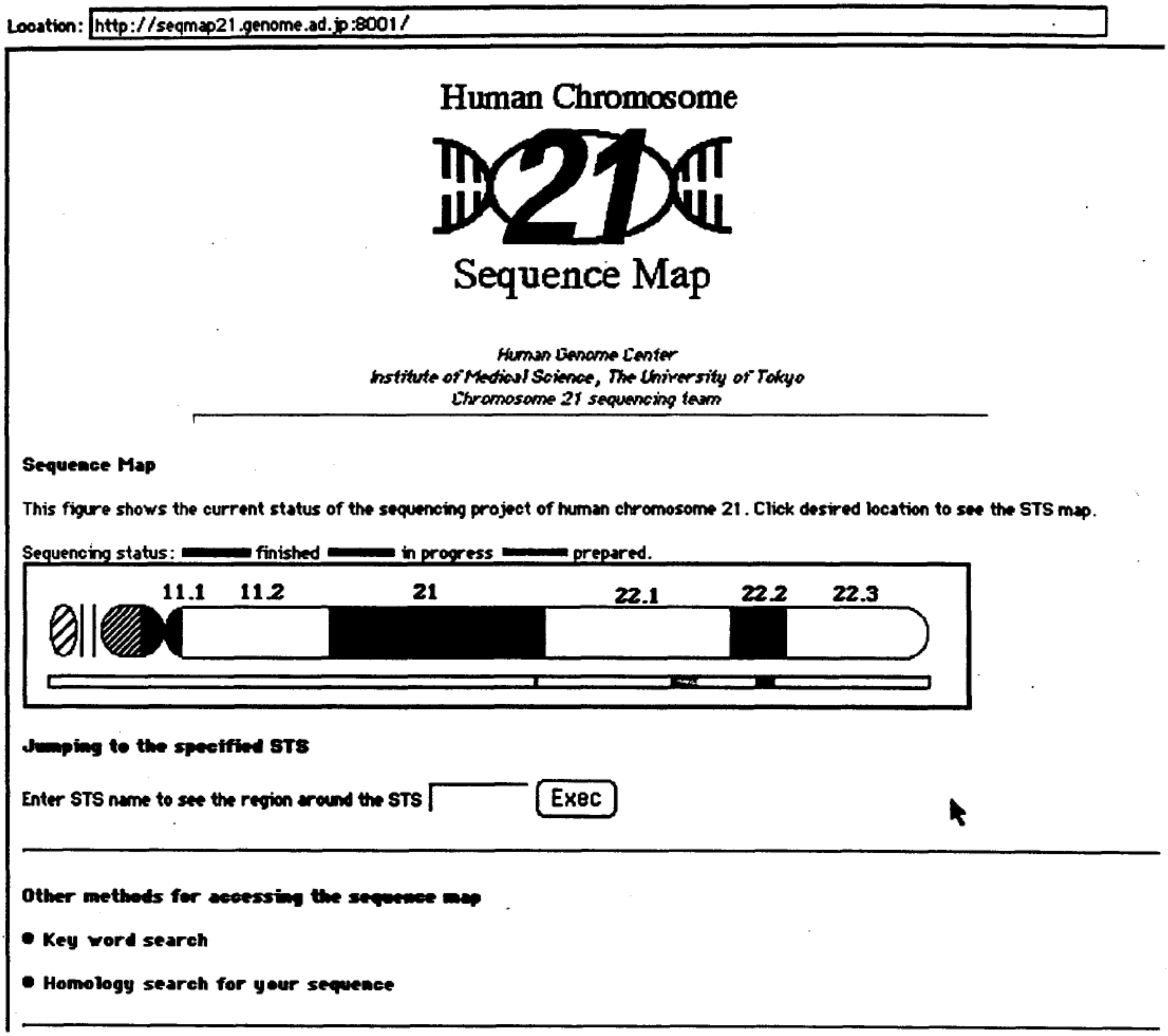
(y;YAC: p;PAC: b; BAC)

? May be second site for STS

The locus is defined in cytogenetic bands; the clones to be sequenced are localized in an interval defined by "MB index on the complete X map (Nagaraja et al., 1977), and by bracketing STS markers; "status" indicates the degree of completion of the project, given either as growing contigs with one or a few remaining gaps, in final checking or completed.

All sequencing are done by OSS approach, except those labeled with * (which were done by random shotgun sequencing).

ASAO
FUJIYAMA
(JAPANESE
PROGRAM)



Location: http://seqmap21.genome.ad.jp:8001/

Human Chromosome

21

Sequence Map

Human Genome Center
Institute of Medical Science, The University of Tokyo
Chromosome 21 sequencing team

Sequence Map

This figure shows the current status of the sequencing project of human chromosome 21. Click desired location to see the STS map.

Sequencing status: ▬ finished ▬ in progress ▬ prepared.

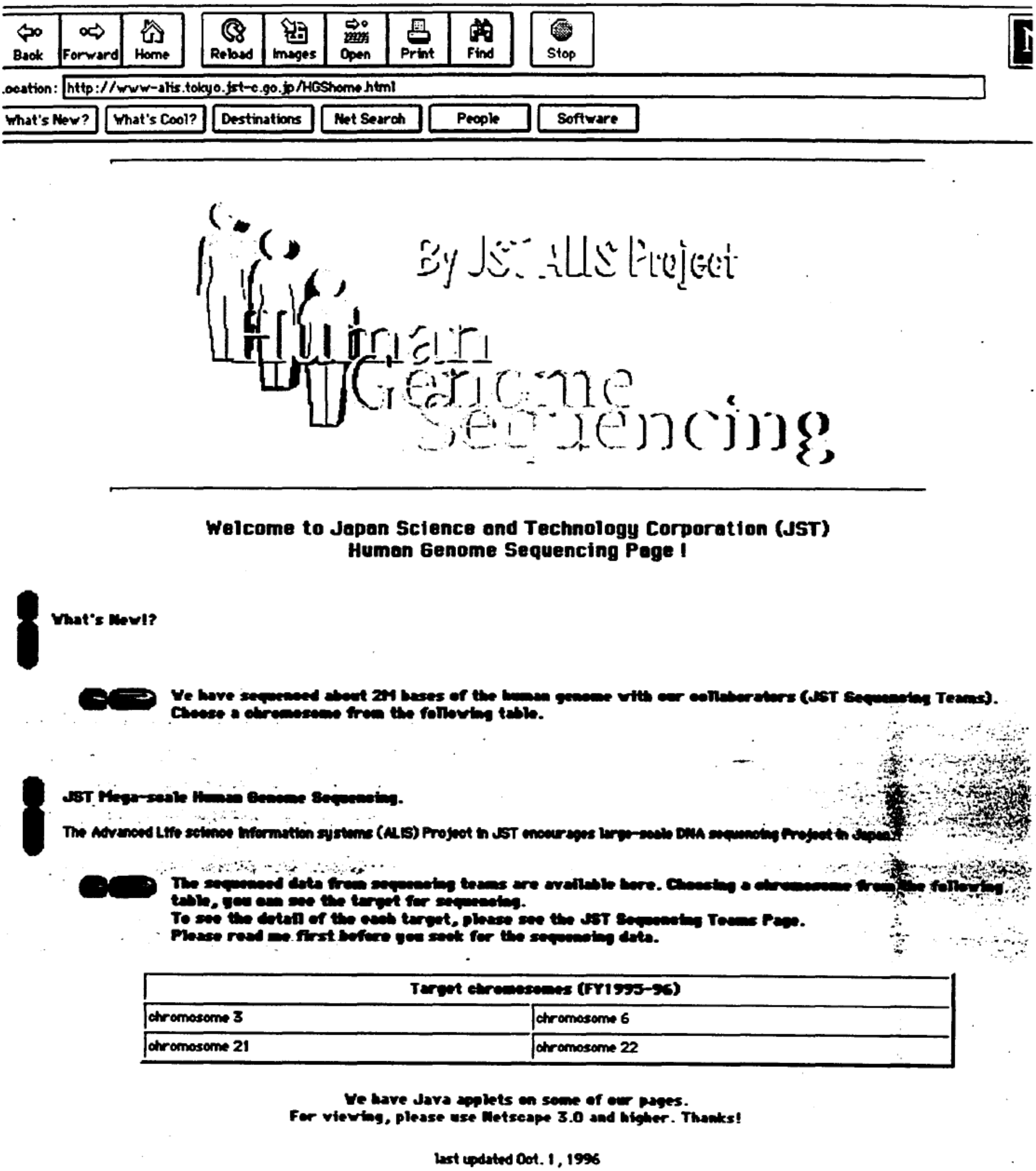
Jumping to the specified STS

Enter STS name to see the region around the STS

Other methods for accessing the sequence map

- Key word search
- Homology search for your sequence

Japan Science and Technology Corporation



By JST ALIS Project

Human Genome Sequencing

Welcome to Japan Science and Technology Corporation (JST)
Human Genome Sequencing Page !

What's New!?

We have sequenced about 2M bases of the human genome with our collaborators (JST Sequencing Teams). Choose a chromosome from the following table.

JST Mega-scale Human Genome Sequencing.

The Advanced Life science Information systems (ALIS) Project in JST encourages large-scale DNA sequencing Project in Japan.

The sequenced data from sequencing teams are available here. Choosing a chromosome from the following table, you can see the target for sequencing.
To see the detail of the each target, please see the JST Sequencing Teams Page.
Please read me first before you seek for the sequencing data.

Target chromosomes (FY1995-96)	
chromosome 3	chromosome 6
chromosome 21	chromosome 22

We have Java applets on some of our pages.
For viewing, please use Netscape 3.0 and higher. Thanks!

last updated Oct. 1, 1996

● Sequencing Schedule

Target¥FY	1995	1996	1997	1998	Total
3p21.3	1,000kbp	-	-	-	1,000kbp
8p11.2	300kbp	1,000kbp	1,200kbp	-	2,500kbp
8p21.3-p22	-	-	200kbp	800kbp	1,000kbp
9q32	-	700kbp	300kbp	-	1,000kbp
Total	1,300kbp	1,700kbp	1,700kbp	800kbp	5,500kbp

This plan may be altered by annual budgeting.

● Sequencing Schedule

Target¥FY	1995	1996	1997	1998	Total
6p21.3	150kbp	400kbp	450kbp	200kbp	1,200kbp
Total	150kbp	400kbp	450kbp	200kbp	1,200kbp

This plan may be altered by annual budgeting.

● Sequencing Schedule

Target¥FY	1995	1996	1997	1998	Total
21q22.2	400kbp	800kbp	-	-	1,200kbp
21q22.1	-	1,000kbp	1,000kbp	-	2,000kbp
21q22.3	-	500kbp	2,000kbp	2,000kbp	4,500kbp
Total	400kbp	2,300kbp	3,000kbp	2,000kbp	7,700kbp

This plan may be altered by annual budgeting.

● Sequencing Schedule

Target¥FY	1995	1996	1997	1998	Total
21q22.2	100kbp	400kbp	-	-	500kbp
21q22.3	-	100kbp	500kbp	300kbp	900kbp
22q11.2	500kbp	800kbp	-	-	1,300kbp
22q11	-	-	800kbp	400kbp	1,200kbp
Total	600kbp	1,300kbp	1,300kbp	700kbp	3,900kbp

This plan may be altered by annual budgeting.

年次計画と所要経費

年　度	98 (H10)	99 (H11)	2000 (H12)	01 (H13)	02 (H14)	03 (H15)	04 (H16)	05 (H17)
データ生産能力	15Mb	30Mb	60Mb	60Mb	60Mb	60Mb	60Mb	60Mb
解析対象	h21	h21／m21	m21／h11	h11	h11／m11	m11	h／m	h／m
人員* リソース	6(4)人	12(8)	15(12)	15(12)	15(12)	15(12)	15(12)	15(12)
シークエンス	12(10)人	24(20)	40(36)	40(36)	40(36)	40(36)	40(36)	40(36)
データ処理	4(3)人	5(4)	10(8)	10(8)	10(8)	10(8)	10(8)	10(8)
技術開発	1人	3(2)	6(4)	6(4)	6(4)	6(4)	6(4)	6(4)
事務部門	2(1)人	4(2)	8(4)	8(4)	8(4)	8(4)	8(4)	8(4)
計	26(18)人	48(36)	79(64)	79(64)	79(64)	79(64)	79(64)	79(64)
運営経費	20億円	30億円	60億円	60億円	60億円	60億円	60億円	60億円

*（　）は人材派遣で可な人数

ハード開発　←→　ソフト開発

GLEN EVANS

UTSW Genome Science and Technology Center

Ongoing Projects:

- **NCHGR Genome Science and Technology Center - Sequencing portions of chromosome 11, 15**

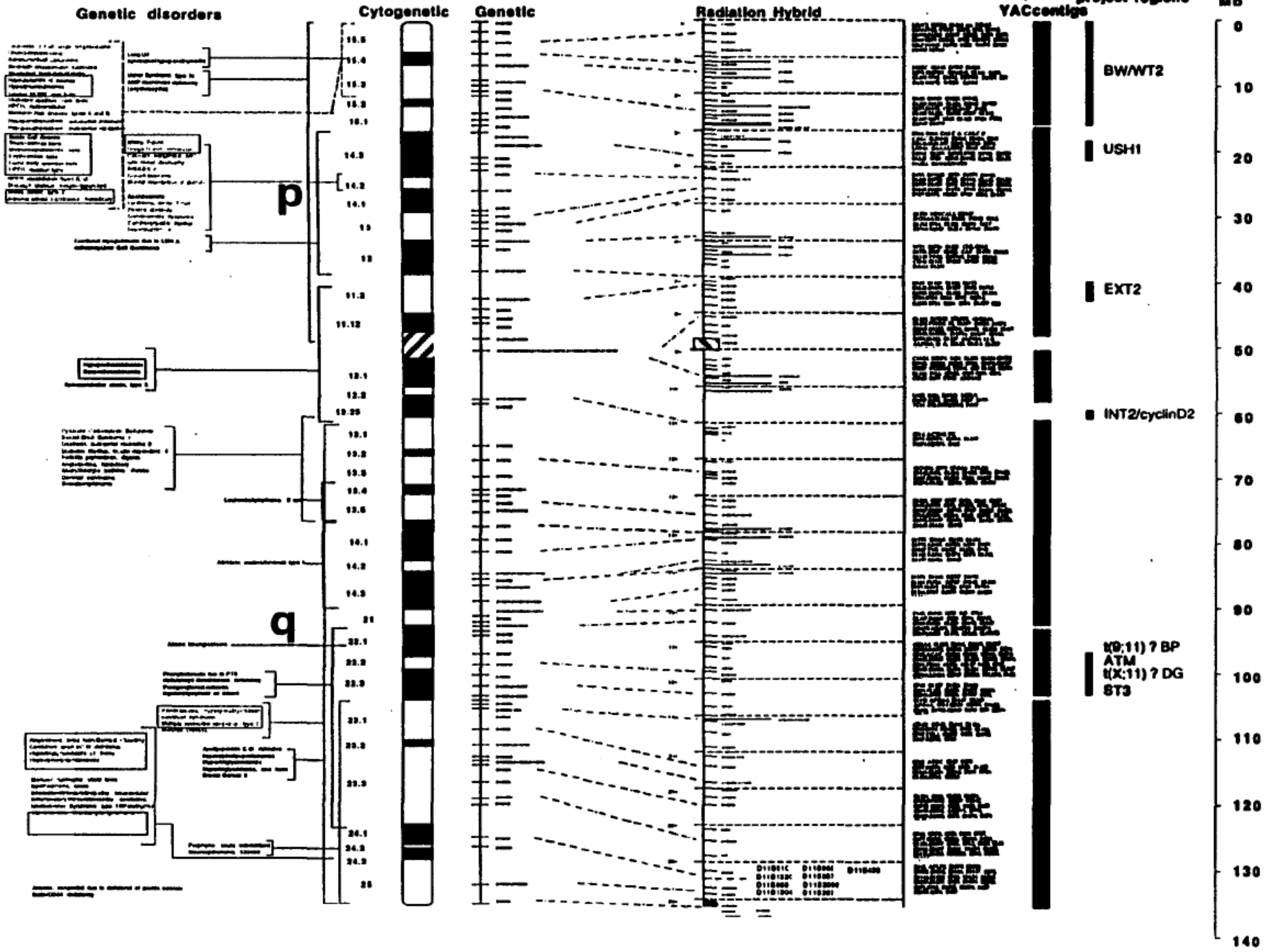
- **Department of Energy - PAC/BAC end-sequence data resource for sequencing the human genome (consortium with RPCI, Cedars-Sinai Medical Center)**

- **Collaborations with Hewlett Packard/Convex, Beckman Instruments/Sagian, Texas Instruments, Nanogen.**

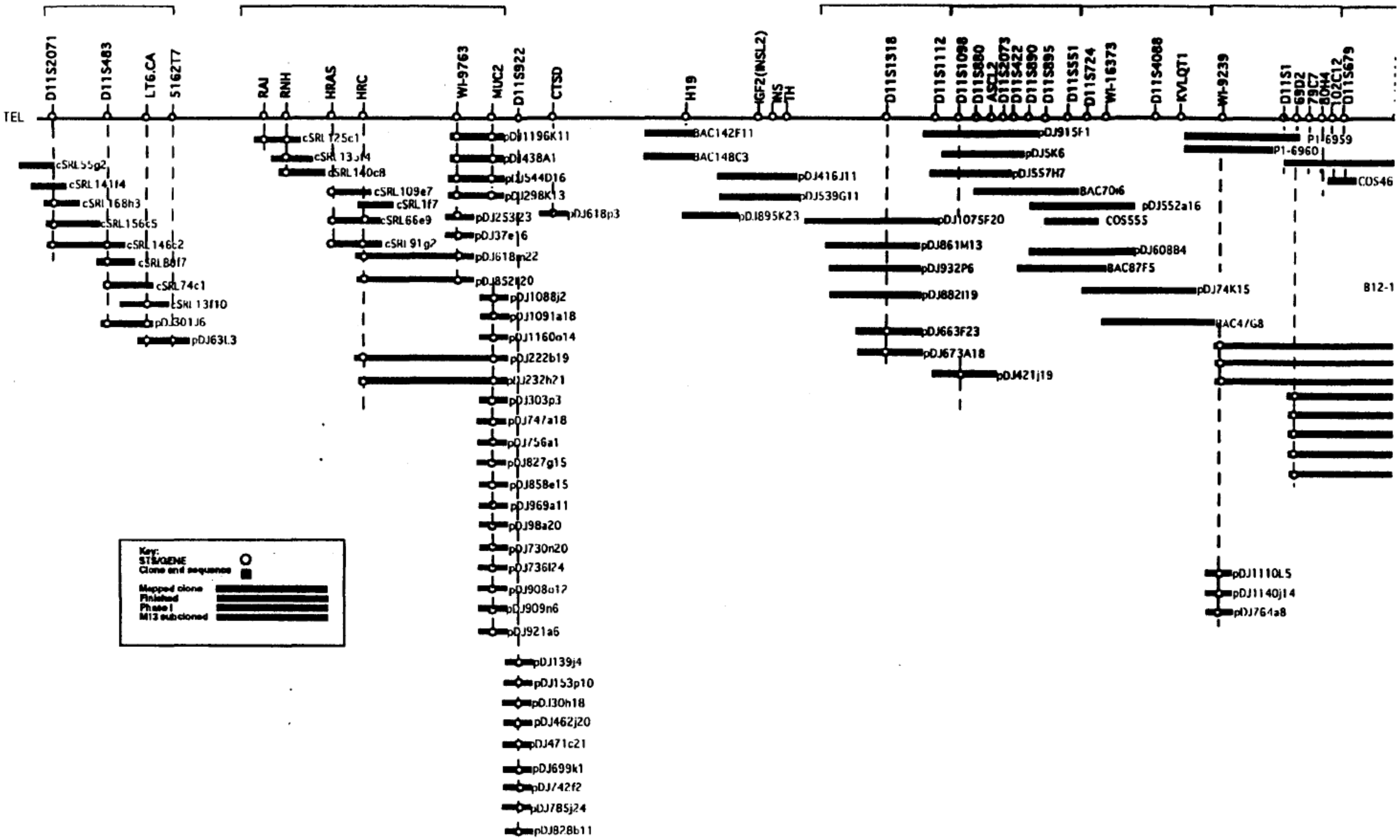
Map Construction

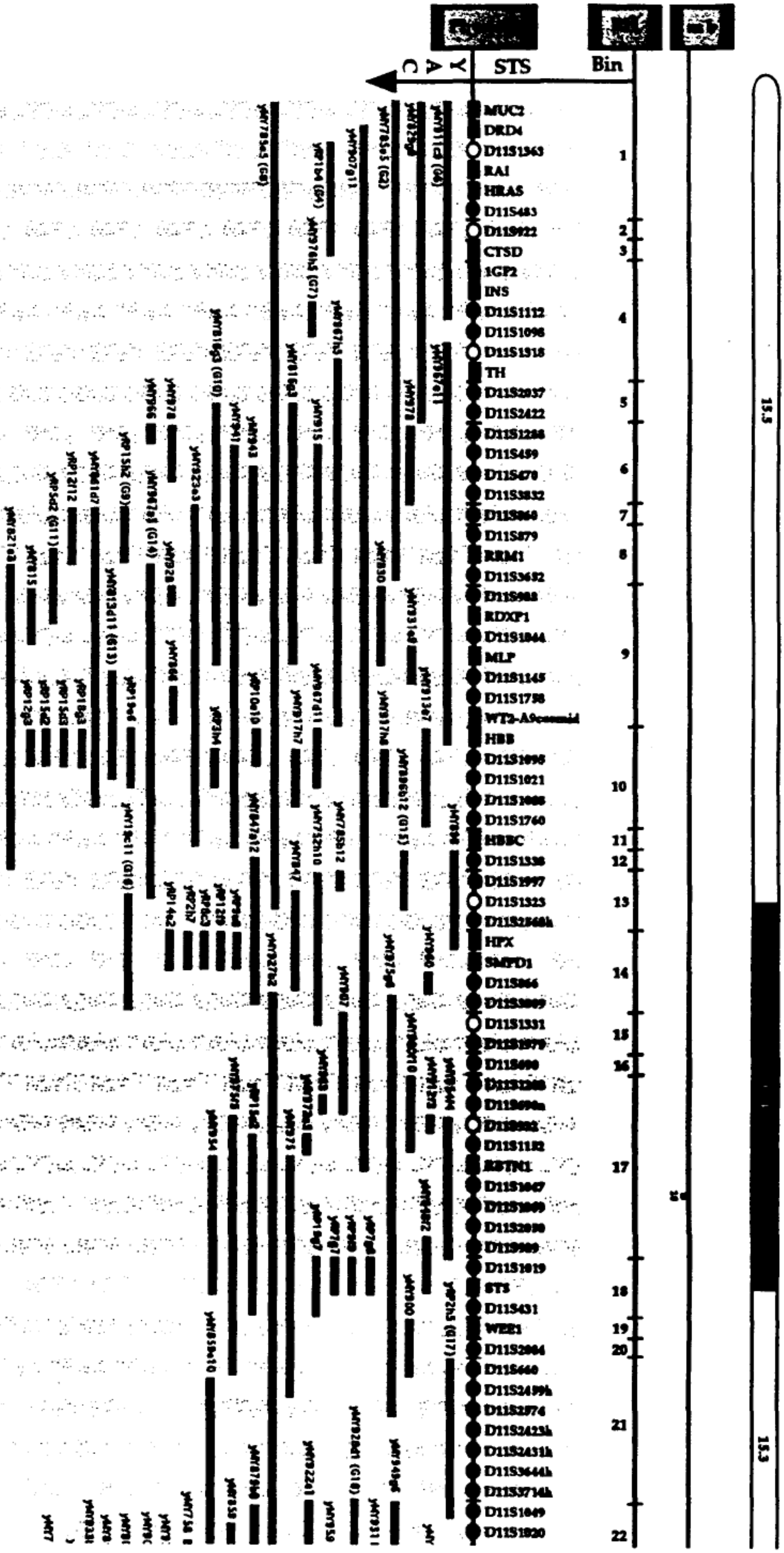
- Based on YAC/STS content map (905 STSs) supplemented with 17,965 "binned" cosmid end-sequences (chr 11), FACS sorted M13 sequences (chr 15)
- Conversion to PAC/BAC map
- PAC/BAC isolation by high density grid hybridization with pooled STS-specific oligonucleotides (20X)
- Confirmation by PCR with STSs (5X)
- Four restriction enzyme fingerprints of each PAC
- PAC/BAC end-sequencing of all clones to detect overlaps, generate additional "gap-filling" STSs and assemble map
- All PACs FISH confirmed to eliminate chimeras (<2%)
- Map becomes the display feature of sequence presentation on WWW

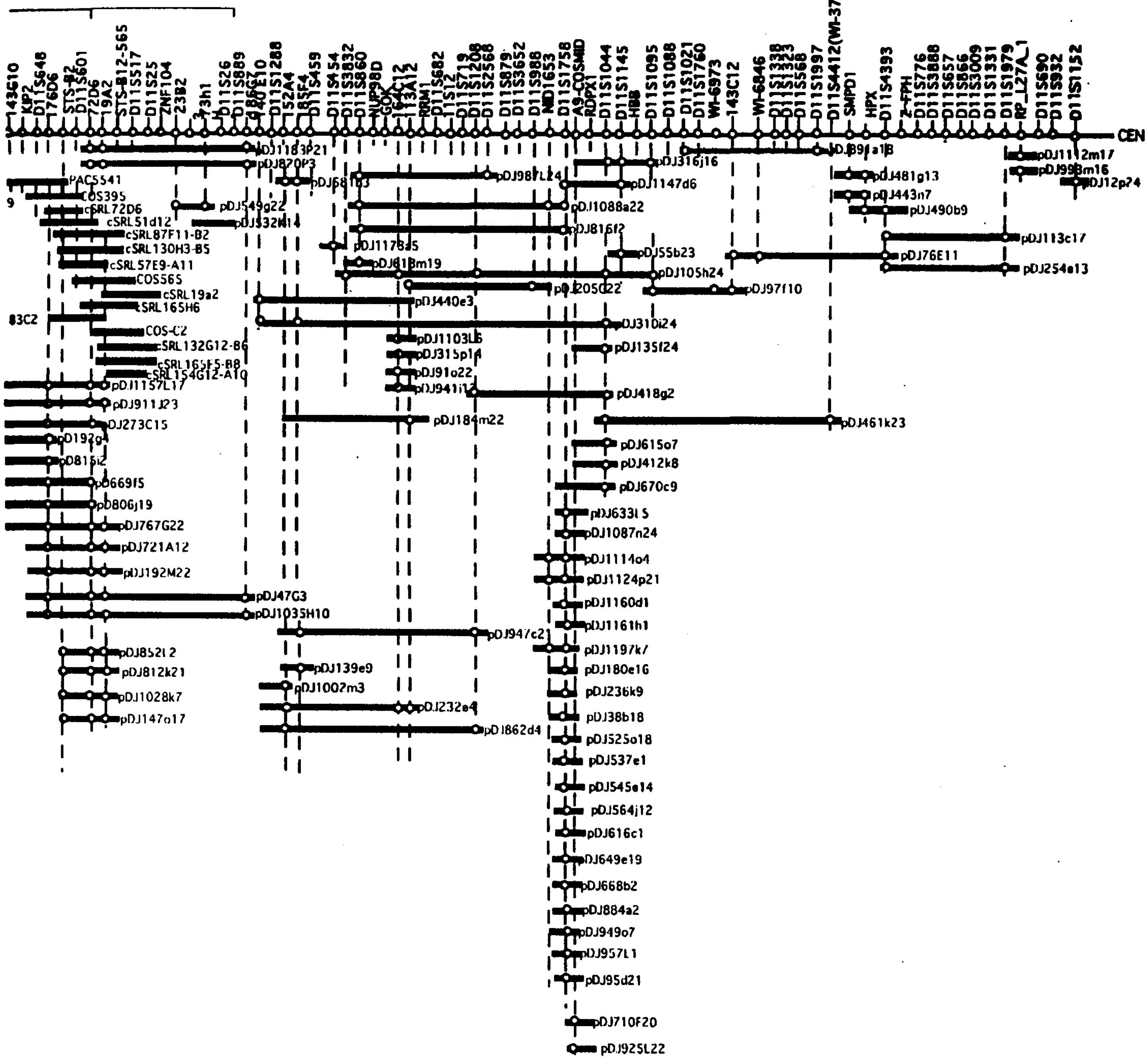
Chromosome 11 Integrated Map



SEQUENCING MAP - CHROMOSOME 11p15.5 - RH BINS 1- 16




YAC
STS
Bin
MUC2
DRD4
D11S1363
1
RAI
HRAS
D11S483
D11S922
2
CTSD
3
IGF2
INS
D11S1112
4
D11S1098
D11S1318
TH
D11S2037
5
D11S2422
D11S1288
D11S459
D11S470
6
D11S3832
D11S860
7
D11S879
RRM1
8
D11S3692
D11S988
RDXP1
D11S1044
MLP
9
D11S1145
D11S1758
WT2-A9cosmid
HBB
D11S1098
D11S1021
10
D11S1088
D11S1760
HBBC
11
D11S1338
12
D11S1997
D11S1323
13
D11S2560h
HPX
SMPD1
14
D11S866
D11S3009
D11S1331
15
D11S1979
D11S890
16
D11S2368
D11S890a
D11S902
D11S1182
17
RBTN1
D11S1047
D11S1809
D11S2030
D11S909
D11S1019
STS
18
D11S431
WEE1
19
D11S2004
20
D11S660
D11S2459h
D11S2574
D11S2423h
21
D11S2431h
D11S3644h
D11S3714h
D11S1049
D11S1020
22
15.5
15.3


143G10
KIP2
D11S648
176D6
STS-B2
D11S601
72D6
19A2
STS-B12-565
D11S517
D11S25
ZNF104
23B2
3
73h1
H
D11S26
D11S889
486G7
40E10
D11S1288
152A4
185F4
D11S459
D11S454
D11S3832
D11S860
NUP98D
GOK
164C12
113A12
RRM1
D11S682
D11S12
D11S719
D11S1208
D11S2568
D11S879
D11S3652
D11S988
MD1653
D11S1758
A9-COSMID
RDPX1
D11S1044
D11S1145
HBB
D11S1095
D11S1088
D11S1021
D11S1760
WI-6973
143C12
WI-6846
D11S1338
D11S1323
D11S568
D11S1997
D11S4412(WI-3787)
SMPD1
HPX
D11S4393
Z-FPH
D11S776
D11S3888
D11S657
D11S866
D11S3009
D11S1331
D11S1979
RP_L27A_1
D11S690
D11S932
D11S1152
CEN
pDJ1183P21
pDJ870P3
PAC5541
9
COS395
cSRL72D6
cSRL51d12
cSRL87F11-B2
cSRL130H3-B5
cSRL57E9-A11
COS565
cSRL19a2
cSRL165H6
83C2
COS-C2
cSRL132G12-B6
cSRL165F5-B8
cSRL154G12-A10
pDJ1157L17
pDJ911J23
pDJ273C15
pDJ192g4
pDJ815i2
pDJ669f5
pDJ806j19
pDJ767G22
pDJ721A12
pDJ192M22
pDJ47G3
pDJ1035H10
pDJ852I2
pDJ812k21
pDJ1028k7
pDJ147o17
pDJ549g22
pDJ532K14
pDJ681b3
pDJ1178a5
pDJ818m19
pDJ440e3
pDJ184m22
pDJ947c21
pDJ139e9
pDJ1007m3
pDJ232e4
pDJ862d4
pDJ987L24
pDJ1088a22
pDJ816f2
pDJ105h24
pDJ205G22
pDJ310i24
pDJ1103L6
pDJ315p14
pDJ91o22
pDJ941i17
pDJ418g2
pDJ135f24
pDJ316j16
pDJ1147d6
pDJ55b23
pDJ97f10
pDJ461k23
pDJ615o7
pDJ412k8
pDJ670c9
pDJ633L5
pDJ1087n24
pDJ1114o4
pDJ1124p21
pDJ1160d1
pDJ1161h1
pDJ1197k7
pDJ180e16
pDJ236k9
pDJ38b18
pDJ525o18
pDJ537e1
pDJ545e14
pDJ564j12
pDJ616c1
pDJ649e19
pDJ668b2
pDJ884a2
pDJ949o7
pDJ957L1
pDJ95d21
pDJ710F20
pDJ925L22
pDJ891a18
pDJ481g13
pDJ443n7
pDJ490b9
pDJ76E11
pDJ113c17
pDJ254e13
pDJ1112m17
pDJ998m16
pDJ12p24

UTSW GESTEC
Map Production

STSs screened (RH bins 1-85)	465
PACs isolated by hyb	3,185
“Hit” rate (av/range)	12.45 (2.5-24.4)
PACs confirmed by PCR	467
Clones fingerprinted	467

UTSW GESTEC
Resource Lab FISH analysis

PACs analyzed by FISH	216
unique signal	213
chimeric signal	3
% putative chimeras	1.3%
band assignment	192
band analysis	142

Sequencing Strategy

○ **M13/plasmid shotgun library of entire PAC < 6X coverage.**

○ **Automated reaction assembly using Sagian/Beckman robot, currently 3,000/day with capacity of 24,000/day.**

○ **Initial 75%primer/25% terminator chemistry and automated assembly using Phred/Phrap.**

○ **Automated synthesis of oligonucleotide primers from initial assembly using Primo software and MerMade 192-channel synthesizers (300/day) for closing and accuracy improvement.**

○ **Finishing using alternate strand reads, long reads, oligo gap closing "auto-finishing" and primer production using Primo, etc.**

○ **Accuracy assessment and additional reads to generate average Phrap score of >40 over entire sequence.**

Sequence levels and estimated accuracy

○ Phase I — Assembled contigs > 1 kb, unordered

○ [illegible]

○ Phase III — Closed contig, no gaps, no resequencing for accuracy improvement, estimated accuracy 10^{-3} to 10^{-4} Genbank acceptable

○ Phase IV — QualPlot analyzed, accuracy improved by resequencing to 10^{-4} based on average Phred/Phrap score > 40

DNA Sequence Production

Level	Type	No.	bp
Data collection	raw data	18	2,160,000
Phase I/II	contigs	41	2,902,496
Phase III	closed	33	1,137,005
Phase IV	$<10^{-4}$ accuracy	4	482,752
Genbank	closed + 10^{-4}	34	1,619,757
Largest contig			341,110

Clone End-Sequencing Project

End-sequence files generated:

Chromosome 11 cosmids	17,965
Giardia lamblia cosmids	2,590
Chromosome 11 PACs	546
Whole Human Genome PACs	1,523

End-sequence database of 5,636,750 bp

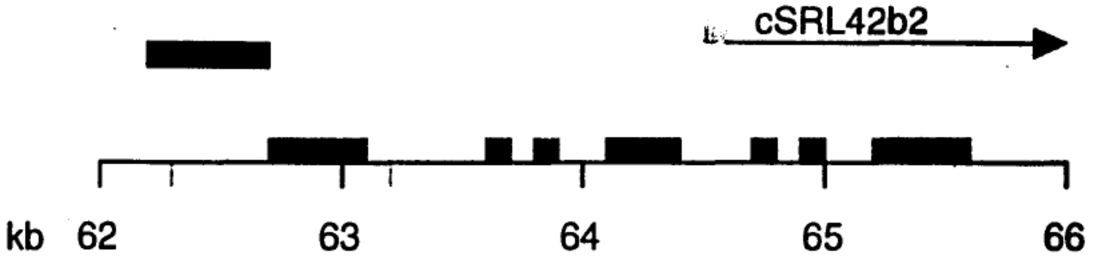
Annotation Protocol

- Final assembly and annotation carried out on HP/Convex Exemplar superparallel computer (8 hrs --> 2 hrs --> 20 minutes)

- Sequence annotated for:
  - Genbank matches
  - EST matches
  - STS matches (map confirmation)
  - End-sequence matches (determination of clone overlap)
  - Grail-predicted exons
  - Repetitive sequence
  - Simple sequence repeats
  - Restriction sites (comparison with fingerprint to confirm assembly)
  - Other features

- QualPlot output - accuracy estimation

Automated sequence annotation



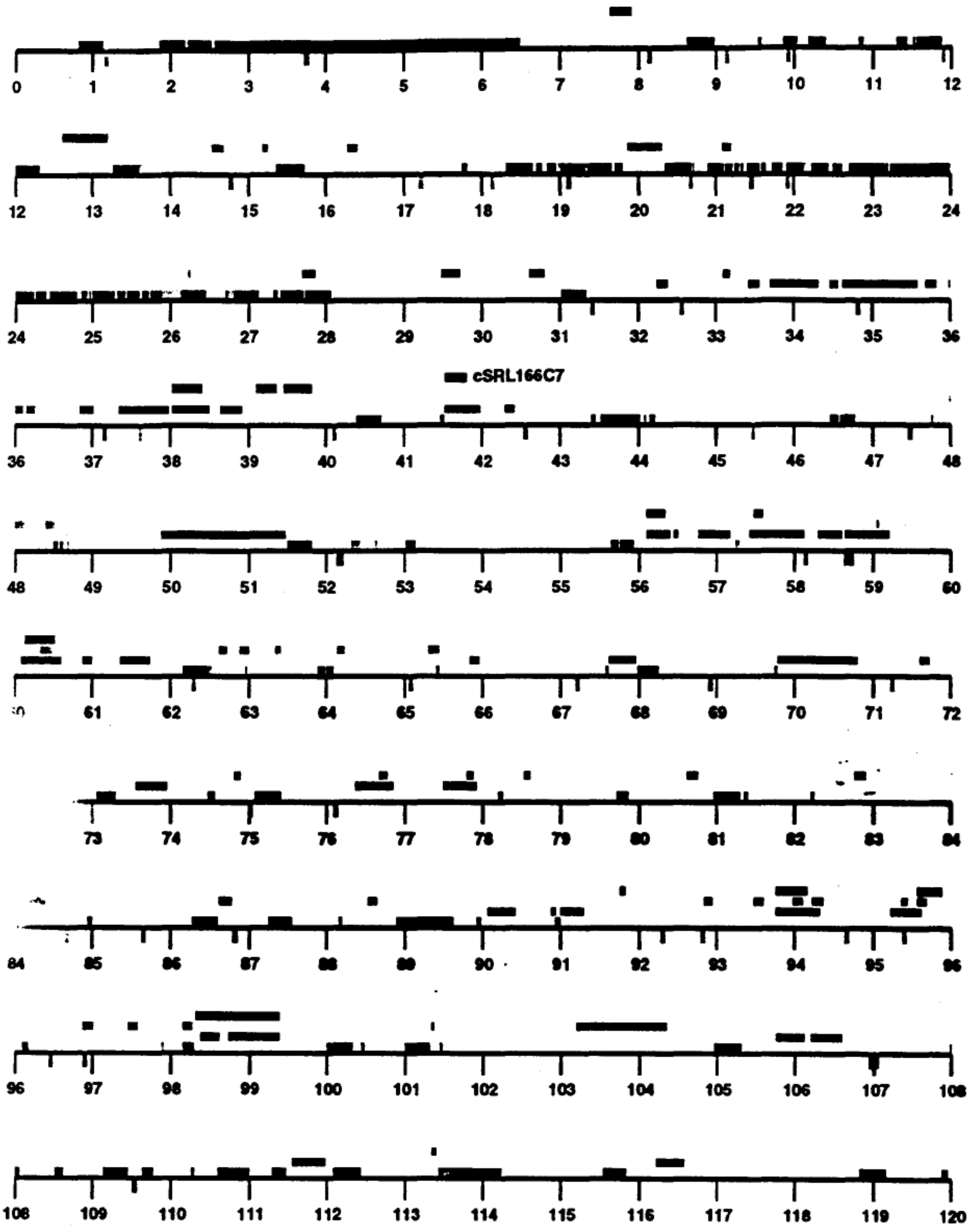
■ Human repetitive element
Simple sequence
■ EST Genbank
■ Non-EST Genbank
. Grail-predicted exon

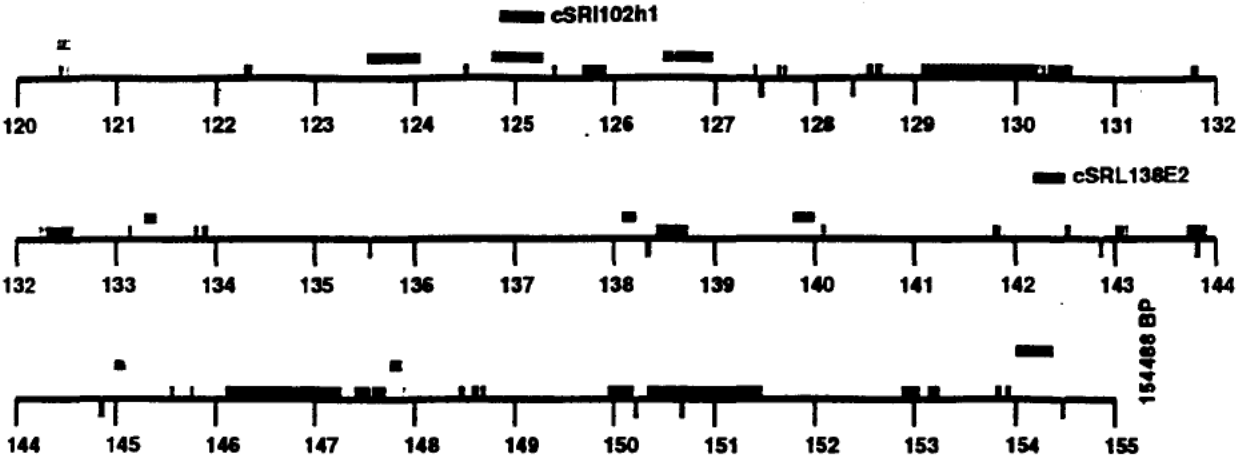
▨ End sequence
(overlapping clone)

| EcoR1, BamH1 sites (QC)

Sequence Features from a 155KB Contig of 11p14.3

Tue Dec 17 08:21:01 CST 1996





LEGEND:
Human Repetitive Element
Simple Sequence
Non-EST GenBank
Predicted Exon
EST GenBank
End Sequence

BamHI: 13.37.67,516,521,1045.1062,2352.2446,2493,2578,3757,4012,4329,4384
4384.4679,5381,5480,5486,5736,5929,6643,8501,9631.9955.10866,13083,15581,17925
17925

EcoRI: 75.470.774.919.979.1092.1188.1233.1987.2252.2259.2493.2542.2777.2798
2798.3804.3851.5321,5363.6097.7183.7299.7361.7948.8564.8993.10062.10654.16818
16818,21332

Data Distribution

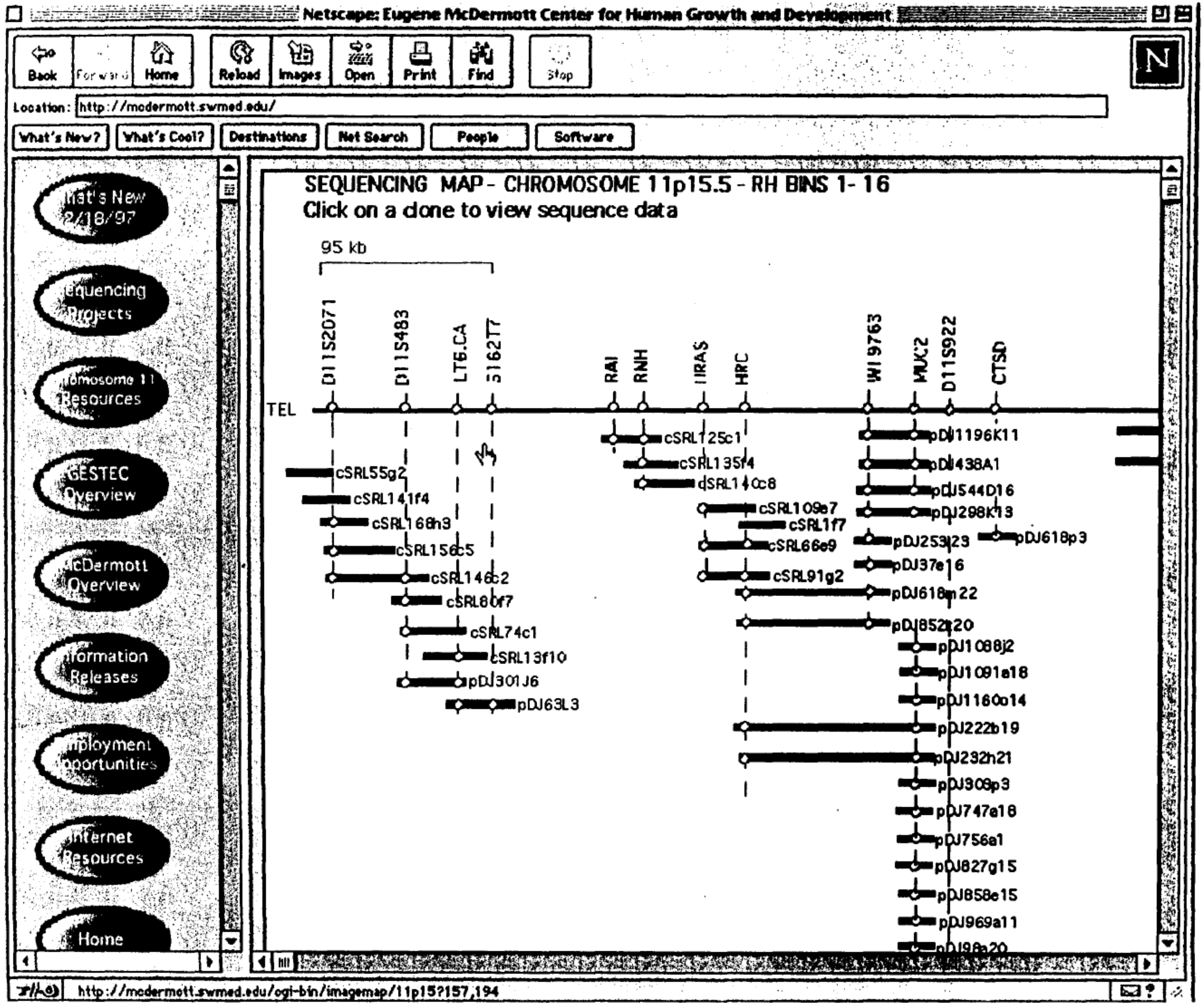
- Maps and sequence available at http//:mcdermott.swmed.edu/ - updated weekly

- Phase I (contigs) and Phase II (closed) made available; unassembled raw data is not made available

- Phase III and Phase IV submitted to Genbank when completed

- WWW display includes:

  Map of sequenced region including clones in progress
  Graphic features display of each clone
  Complete features tables
  QualPlot (accuracy estimate) output



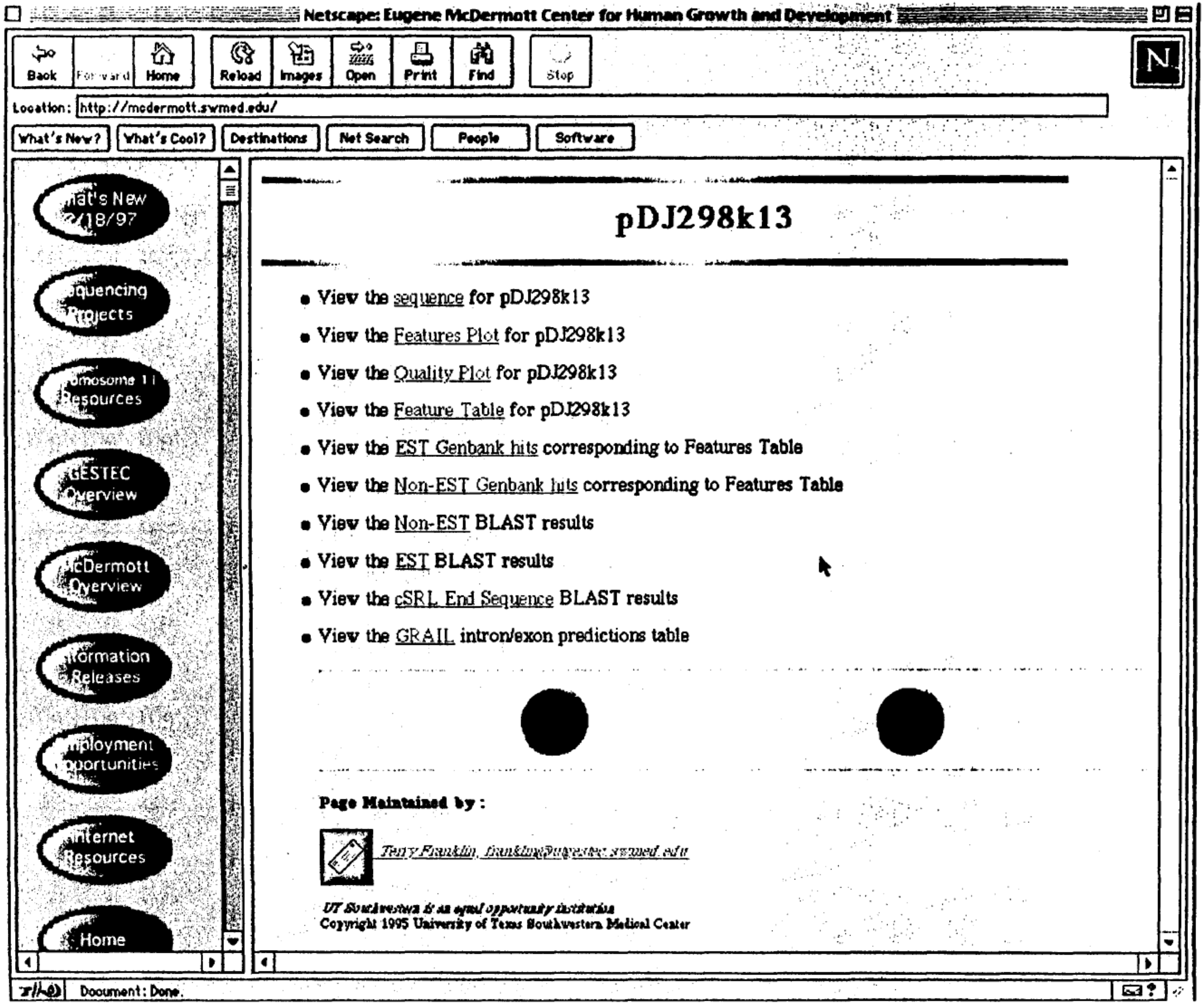
Netscape: Eugene McDermott Center for Human Growth and Development

Location: http://mcdermott.swmed.edu/

SEQUENCING MAP - CHROMOSOME 11p15.5 - RH BINS 1-16

Click on a clone to view sequence data

http://mcdermott.swmed.edu/cgi-bin/imagemap/11p15?157,194



Netscape: Eugene McDermott Center for Human Growth and Development

Location: http://mcdermott.swmed.edu/

pDJ298k13

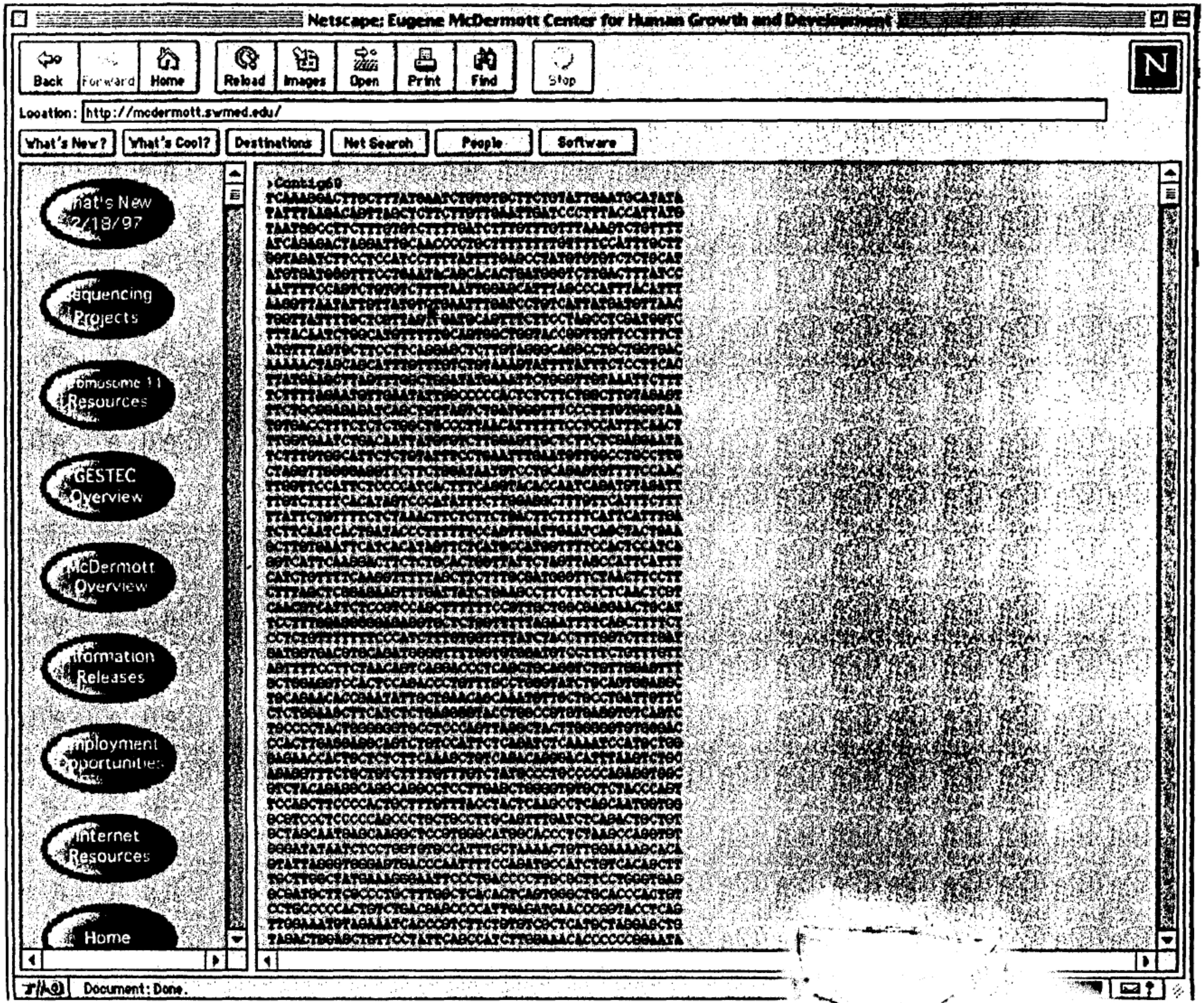
- View the sequence for pDJ298k13
- View the Features Plot for pDJ298k13
- View the Quality Plot for pDJ298k13
- View the Feature Table for pDJ298k13
- View the EST Genbank hits corresponding to Features Table
- View the Non-EST Genbank hits corresponding to Features Table
- View the Non-EST BLAST results
- View the EST BLAST results
- View the cSRL End Sequence BLAST results
- View the GRAIL intron/exon predictions table

Page Maintained by :

Terry Franklin, franklin@genestec.swmed.edu

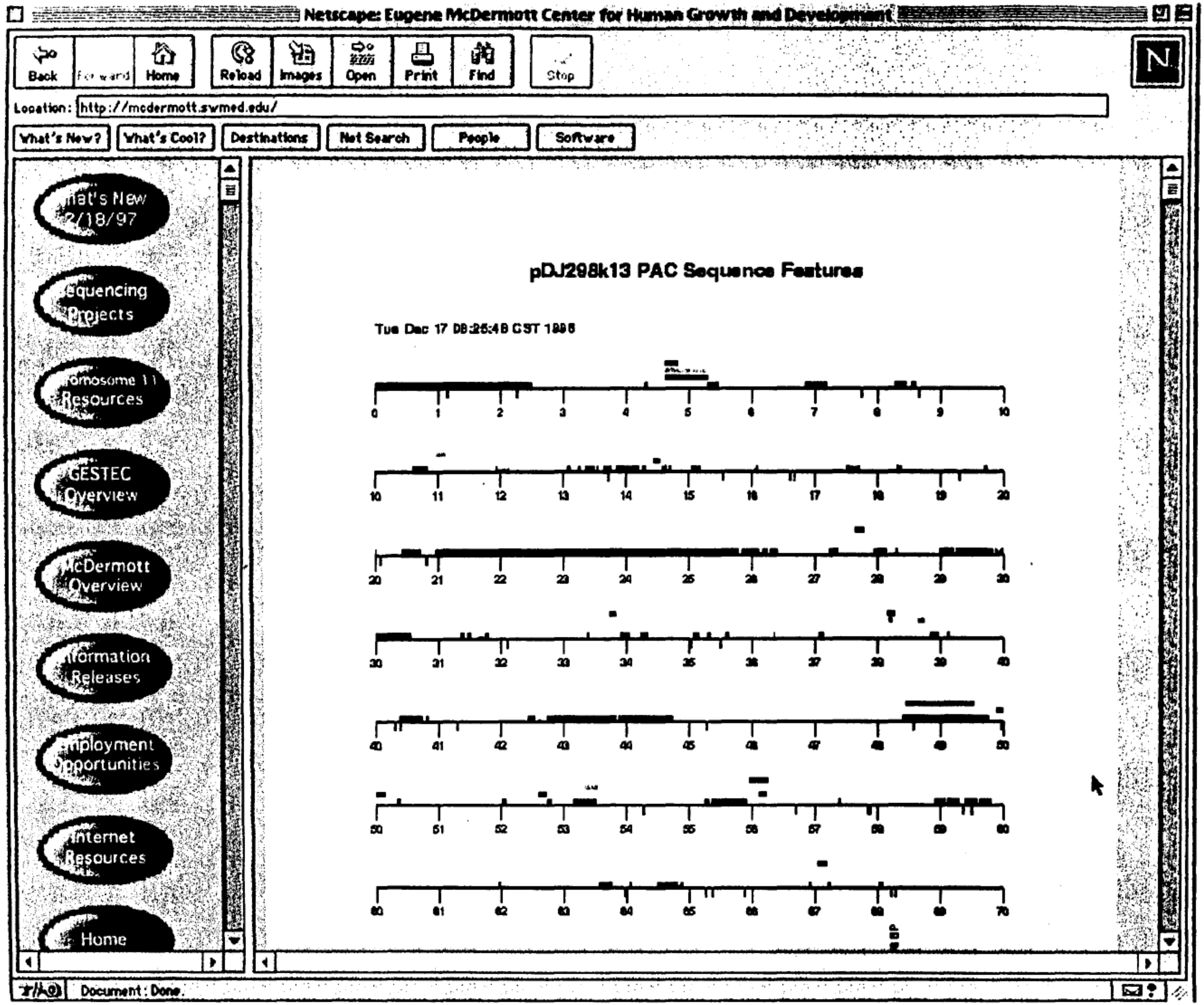
UT Southwestern is an equal opportunity institution

Document: Done.



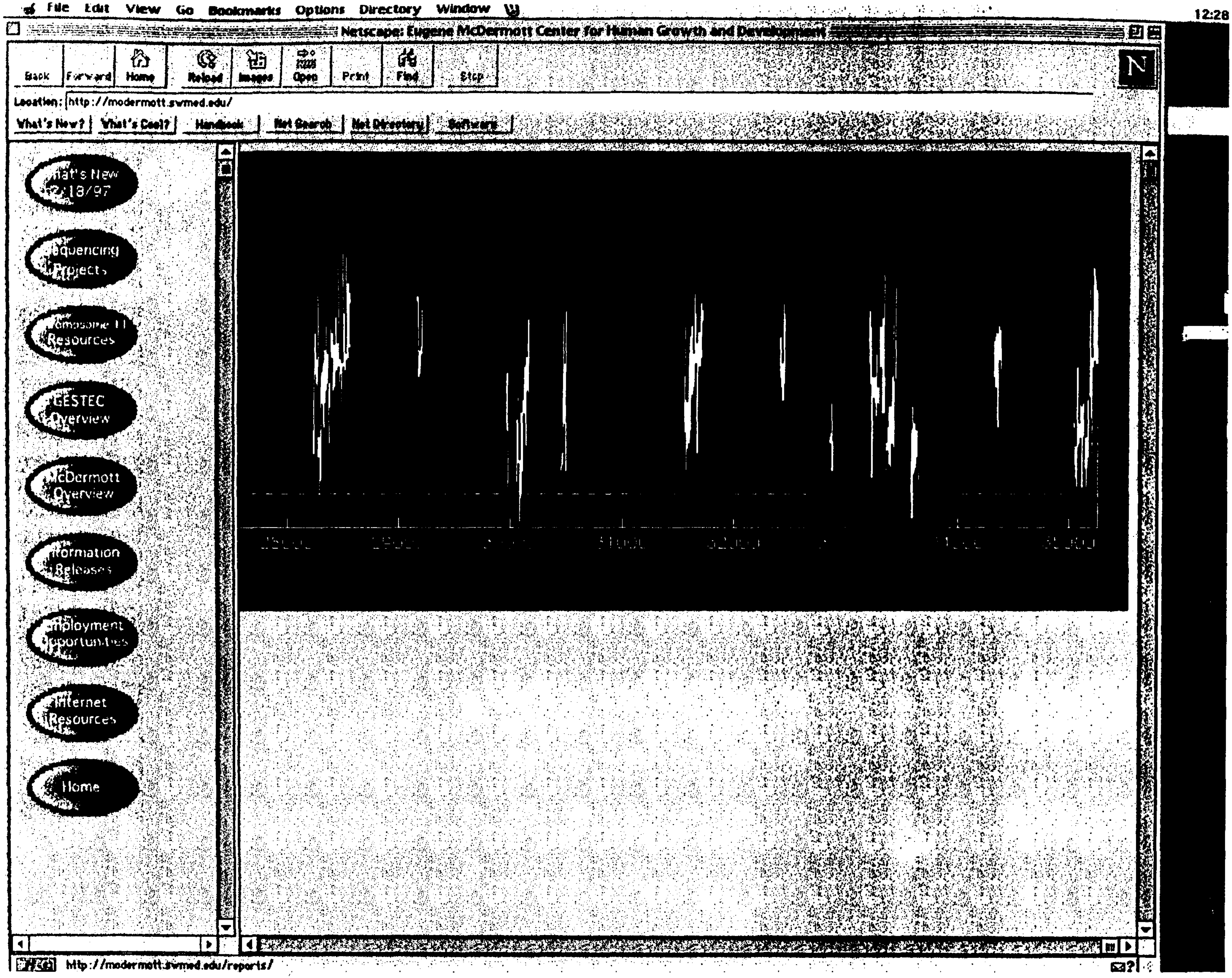
Location: http://mcdermott.swmed.edu/

```
>Contig60
TCAAAGGACTTTGCTTTATGAATCTGTGTGCTTCTGTATTGAATGCATATA
TATTTAAGACAGTTAGCTCTTCTTGTTGAATTGATCCCTTTACCATTATG
TAATGGCCTTCTTTGTGTCTTTTGATCTTTTGTTTGTTTAAAGTCTGTTTT
ATCAGAGACTAGGATTGCAACCCCTGCTTTTTTTTTGTTTTCCATTTGCTT
GGTAGATCTTCCTCCATCCTTTTATTTTGAGCCTATGTGTGTCTCTGCAT
ATGTGATGGGTTTCCTGAATACAGCACACTGATGGGTCTTGACTTTATCC
AATTTTCCAGTCTGTGTCTTTTAATTGGAGCATTTAGCCCATTTACATTT
AAGGTTAATATTGTTATGTGTGAATTTGATCCTGTCATTATGATGTTAAC
[illegible]
```



pDJ298k13 PAC Sequence Features

Tue Dec 17 08:25:48 CST 1996



Netscape: Eugene McDermott Center for Human Growth and Development

Location: http://modermott.swmed.edu/

http://modermott.swmed.edu/reports/

Current Automation Projects

○ MerMade oligonucleotide synthesizer

○ Sagian/Beckman ^{3}S robot

○ Astral DNA sequencer

MerMade

96/192 channel automated oligonucleotide synthesizer

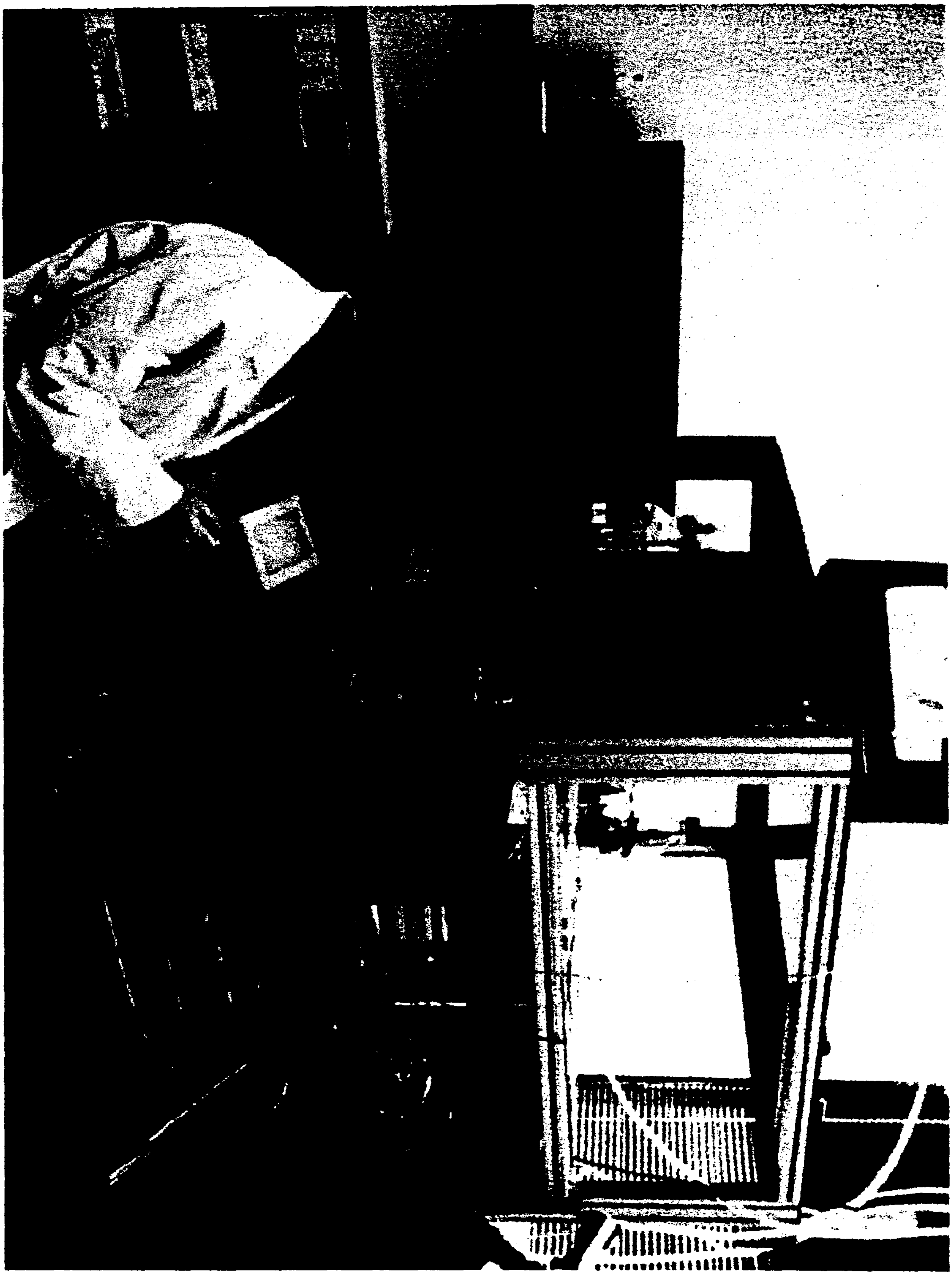
Programmed by Phred/Phrap assemblies using Primo oligonucleotide design software

24 hour/day unattended operation

MerMade I and II in operation, MerMade III constructed, MerMade IV ordered from commercial supplier (Avantec, Inc.)

Cost <$0.10/nucleotide

Available to non-commercial genome centers by no-cost license from University of Texas and contract for construction to Avantec, Inc.



Sagian/Beckman S^3 robot

3 meter rail robot

8 MJ research 96-well PCR thermal cyclers

4 Robbins Hydra 96 channel pipettors

Automated refrigerator storage

Multiple grippers

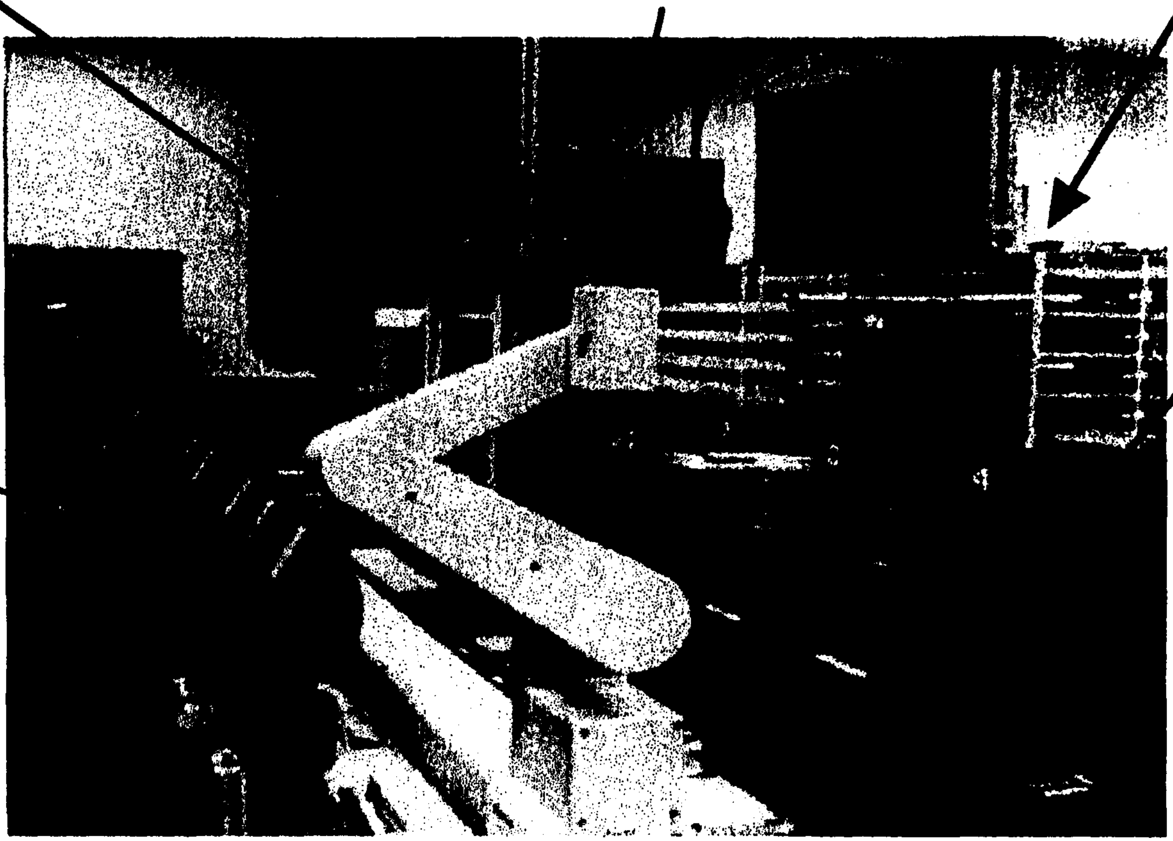
Currently used for all primer/terminator chemistry sequence assemblies in GESTEC

Current capacity 3,000 to 24,000 samples/day

Custom driver software

Developed at UTSW in collaboration with Sagian - available as a commercial product from Beckman Instruments

Sequencing Support System - S3 Production Version



Automated Freezer

ORCA rail robot

Plate hotels

Robbins Scientific 96 channel pipettors (x4)

MJ Research thermal cyclers (x8)

Astral DNA Sequencer

Gel based

High lane density (48, 96, 144, 384)

Multiple dye chemistry without altering hardware
(ABI, ET, Bodipy, other)

Higher sensitivity for dyes

Allows a 5th fluorescent dye for auto lane tracking

Spectral decomposition for increased data quality

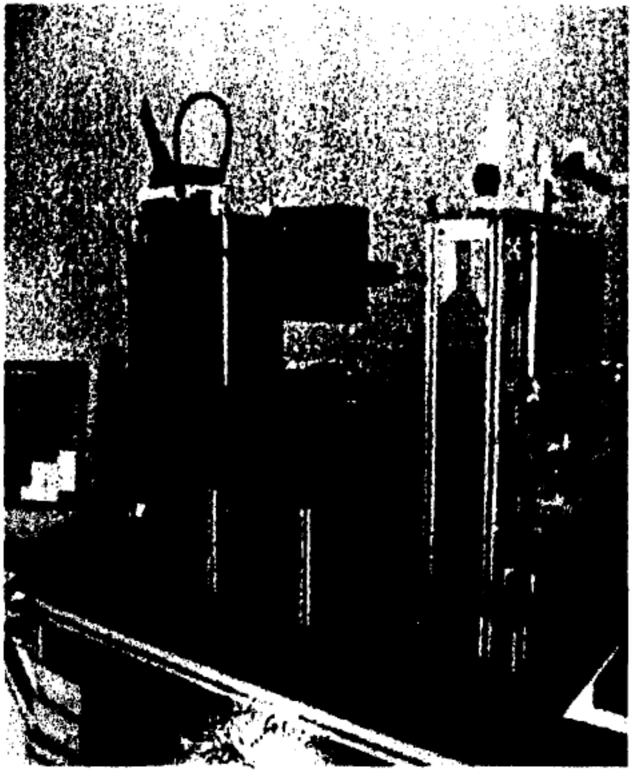
Software for data conversion to industry standard

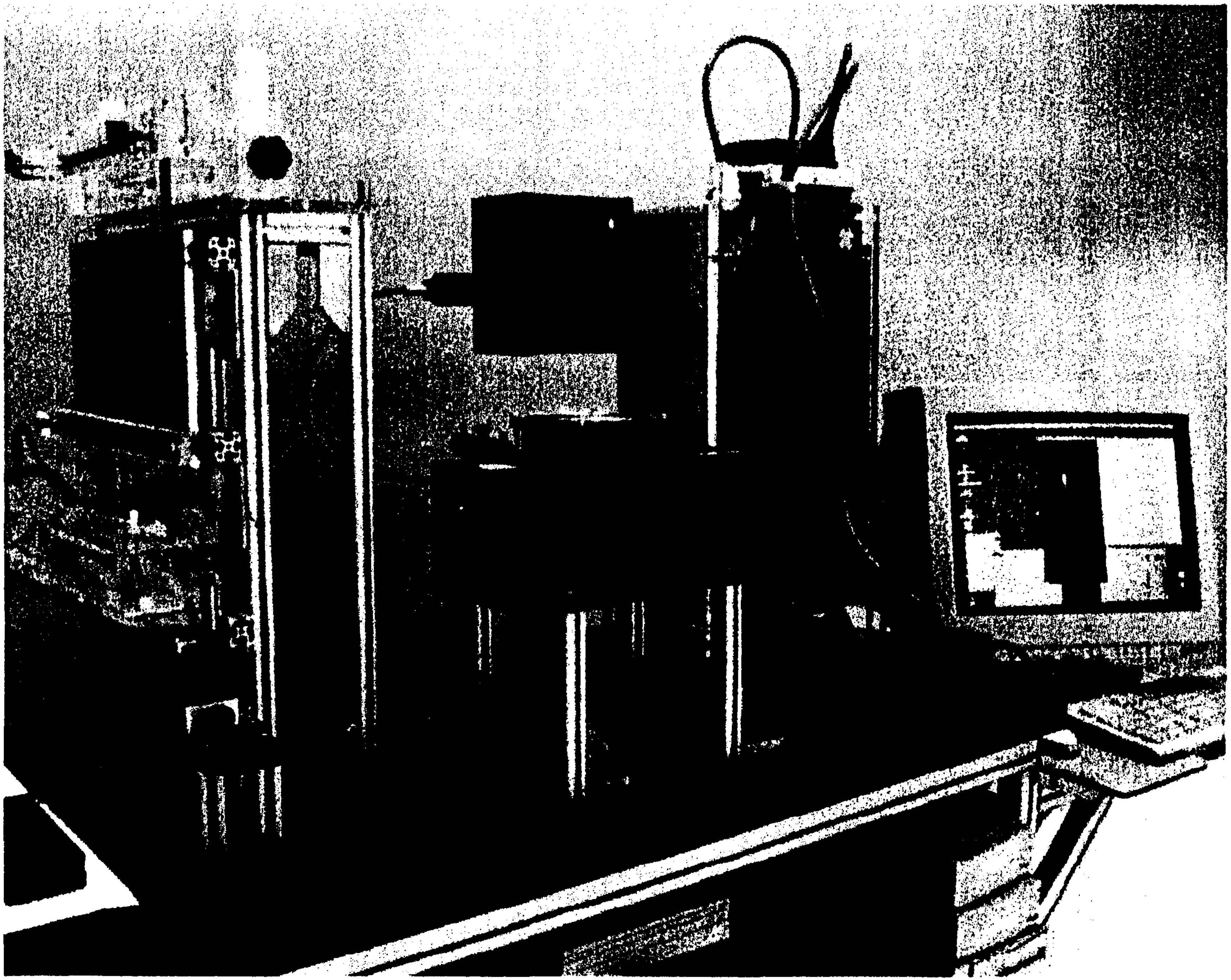
No moving parts for higher reliability

Distribution mechanism under development

ASTRAL, like MISTI employs Hyperspectral Imaging

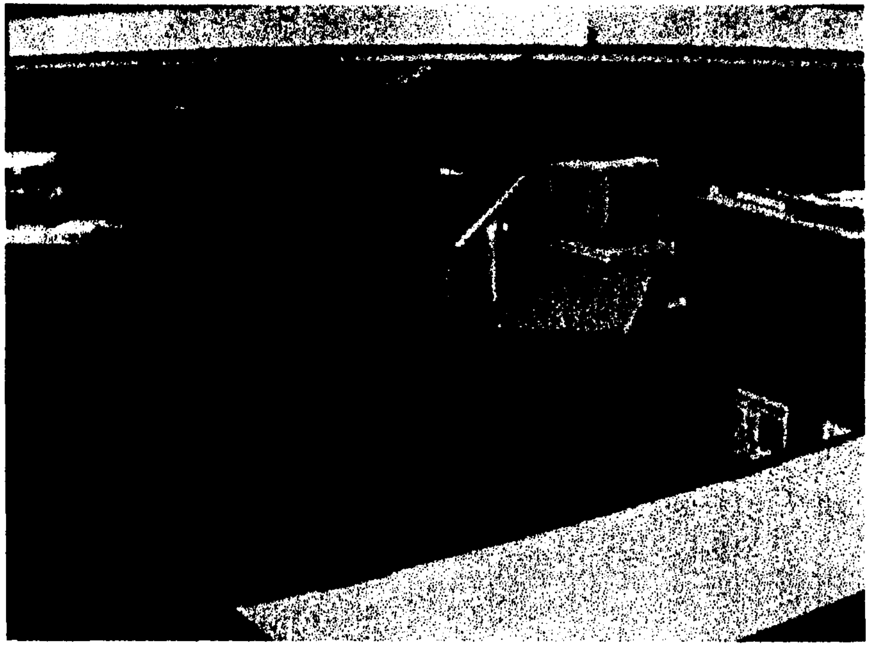
Earth orbit imaging to DNA sequencing





The HP/Convex Parallel Supercomputer

Parallel processing and a large shared memory speed data analysis



8 RISC processers
0.5 GB shared RAM
28 GB Hard Drives
2 Workstations
UNIX, parallelizers
Compilers and optimizers

Groups Sequencing Chromosome 11

Region	Markers		Group	Contact
11p15	11pter	D11S932	UTSW	Evans
11p14	RBTN1	CALCA	GGP	Zabel
11p14	D11S1228	D11S1944E	UTSW	Evans
11p12	D11S1944E	D11S981E	SANGER	Little/Sulston
11p11.2	D11S981E	D11S2399	UTSW	Evans
11q12.2	D11S1368	D11S678	UTSW	Evans
11q13.1	D11S987	D11S3866	UTSW	Evans
11q23	D11S2058	D11S2085	UTSW	Evans

MIKE PALAZZOLO

I. BERKELEY

Drosophila (NIH)

LBNL HGC (DOE)

II. DOE JOINT GENOME INSTITUTE

Livermore

Los Alamos

Berkeley

- Producing Finished Sequence

Totals 5 Mb Drosophila
4 Mb Human

RATE 800 KB / MONTH

IV. STRATEGY

- PHYSICAL MAP
- RANDOM LIGHT SHOTGUN
- BUILD PATHS
- TRANSPOSON-FACILITATED

QUALITY
- ALL DOUBLE STRANDED
- REDUNDANCY FOR ASSEMBLY
- 1 in 10,000

V. SOFTWARE

PATH-BUILDING SUITE
SPACE (ASSEMBLY, ED., MAP)

VI. HARDWARE

COLONY PICKER
LIBRARY POOLING + REPLICATION
OLIGOSYNTHESIZER
AGAROSE GEL IMAGING
AGAROSE GEL LOADER
DNA PREPARATION ROBOT

VII. PARTNERSHIP WITH INDUSTRY AND JGI DOE

GOALS (VOLUME, QUALITY, CYCLE TIME, COST)

PRECISE GOAL DEFINITION

Benchmarking

METRICAL TOOLS

ROADMAPS FOR Operation + TECH.

IMPLEMENTATION + EVOLUTION

NEW SPACE

VIII. METRICAL TOOLS

PROCESS MODEL

COST MODEL

COST ACCOUNTING

PICK - A - MIX

IX. PROCESS MODEL

A. BUILD

SPACE, EQUIPMENT,
LABOR, PROCESS FLOW

B. VALIDATION

PEOPLE AGREE

MODEL OUTPUT MATCHES
FACILITY OUTPUT

PREDICTIVE VALUE

C. UTILITIES

PERSONNEL ACTIVITY RATE
EQUIPMENT UTILIZATION
BOTTLENECK ANALYSIS
LAY OUT ALTERNATIVES
PROCESS ALTERNATIVES

D. BOTTLENECKS

E. PICK-A-MIX /XI CIM

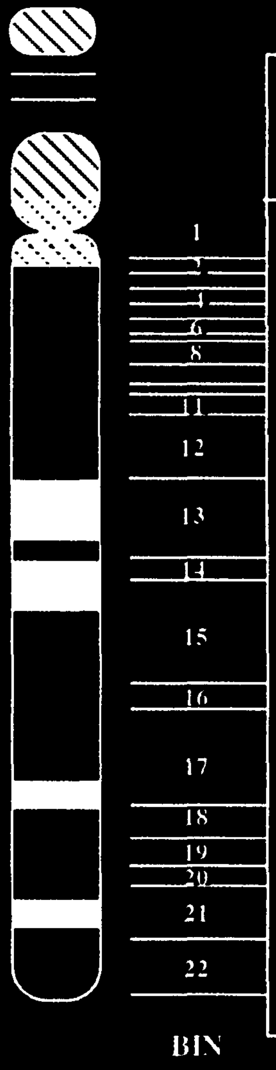
BRUCE ROE

Total sequence data submitted to GenBank

8-31-95-9-1-96	1,846,870 bp
9-2-96 - 11-15-96	1,997,137 bp
Total additional in progress	2,154,832 bp
Total	5,998,839 bp

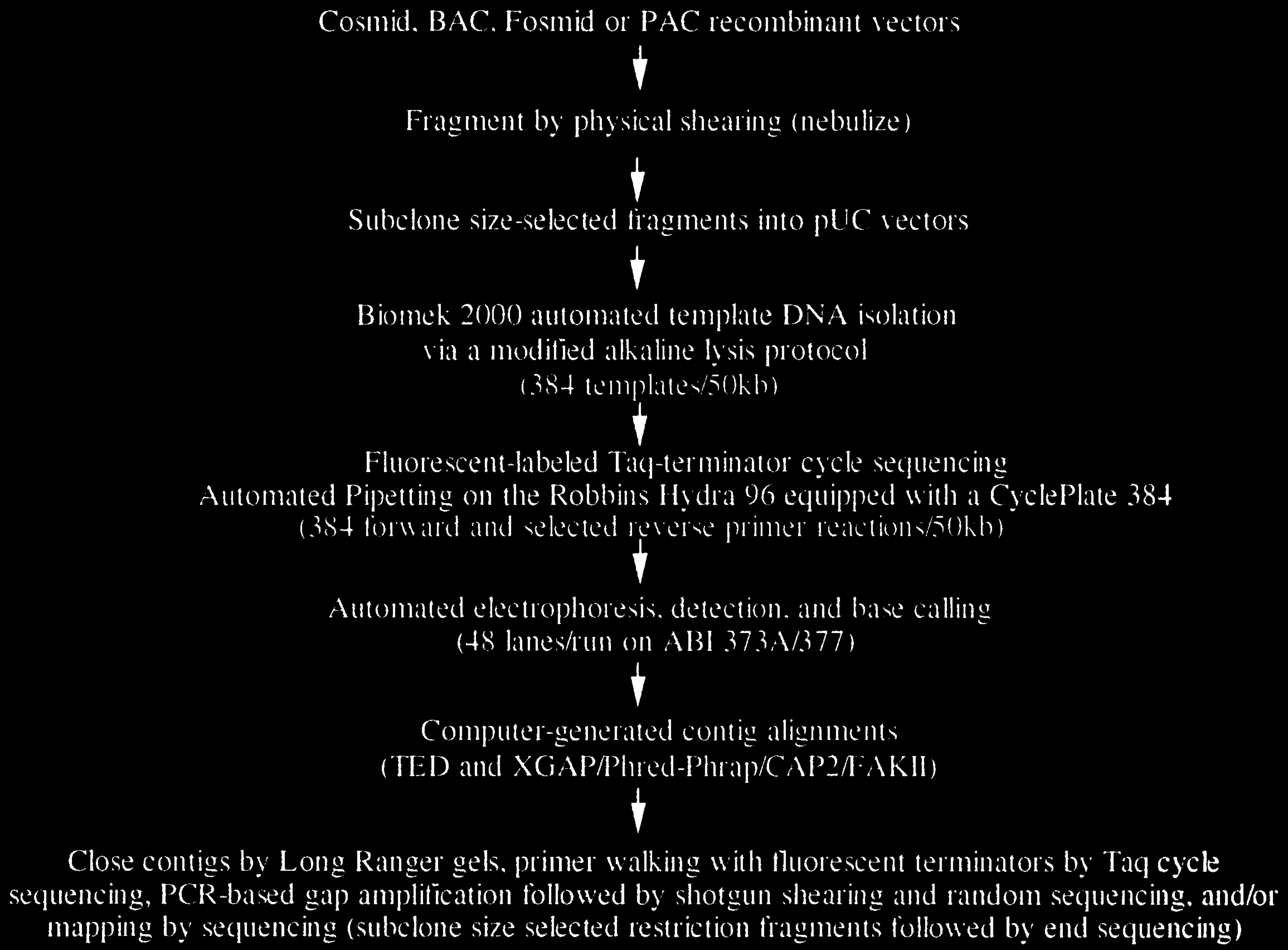
4(2) 377's - Hu/Mo
2(2) 377's - Bact.

OU Chromosome 22 Sequencing Status



Mapped Cosmids, BACs, PACs, Fosmids		Contig Size	Oklahoma Sequencing Status
CER	Oklahoma	1.2 MB	350 Kb 99.9% in GB
DGCR		1.2 MB	in GenBank
		2.0 MB	mapped - CHOP
IGLC		1.1 Mb	480 Kb in GenBank
BCR		152 Kb	in GenBank
		4.0 MB	mapped - CHOP
MDR	WashU and Sanger Center	100 Kb	in GenBank
ES MDR NF2		540 Kb	in GenBank
D22S16		38 Kb	in GenBank
MDR		180 Kb	in GenBank
ACR		140 Kb	in GenBank

Shotgun Cloning, Automated DNA Isolation, Fluorescent-Based DNA Sequencing, and Closure



Cosmid, BAC, Fosmid or PAC recombinant vectors

Fragment by physical shearing (nebulize)

Subclone size-selected fragments into pUC vectors

Biomek 2000 automated template DNA isolation
via a modified alkaline lysis protocol
(384 templates/50kb)

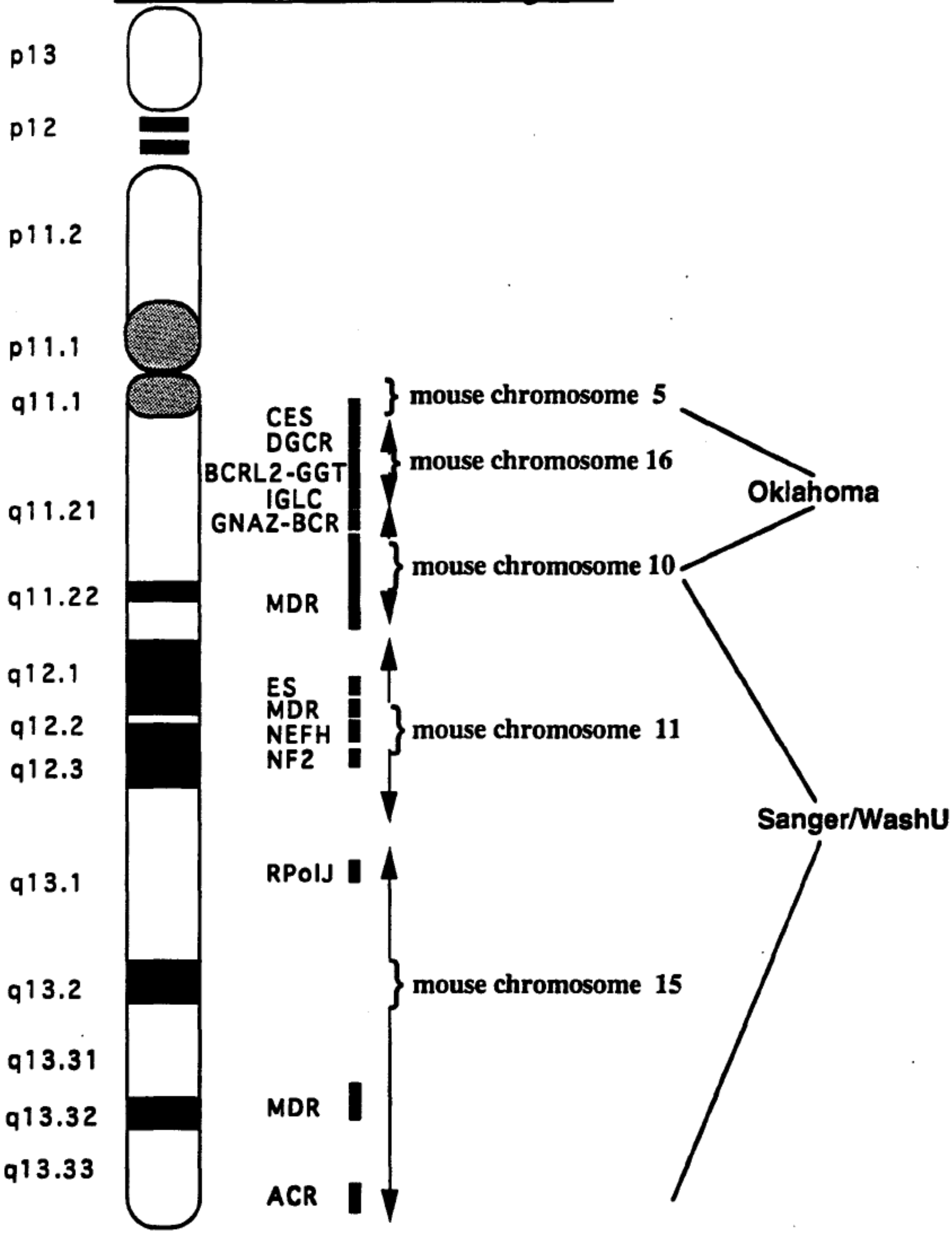
Fluorescent-labeled Taq-terminator cycle sequencing
Automated Pipetting on the Robbins Hydra 96 equipped with a CyclePlate 384
(384 forward and selected reverse primer reactions/50kb)

Automated electrophoresis, detection, and base calling
(48 lanes/run on ABI 373A/377)

Computer-generated contig alignments
(TED and XGAP/Phred-Phrap/CAP2/FAKII)

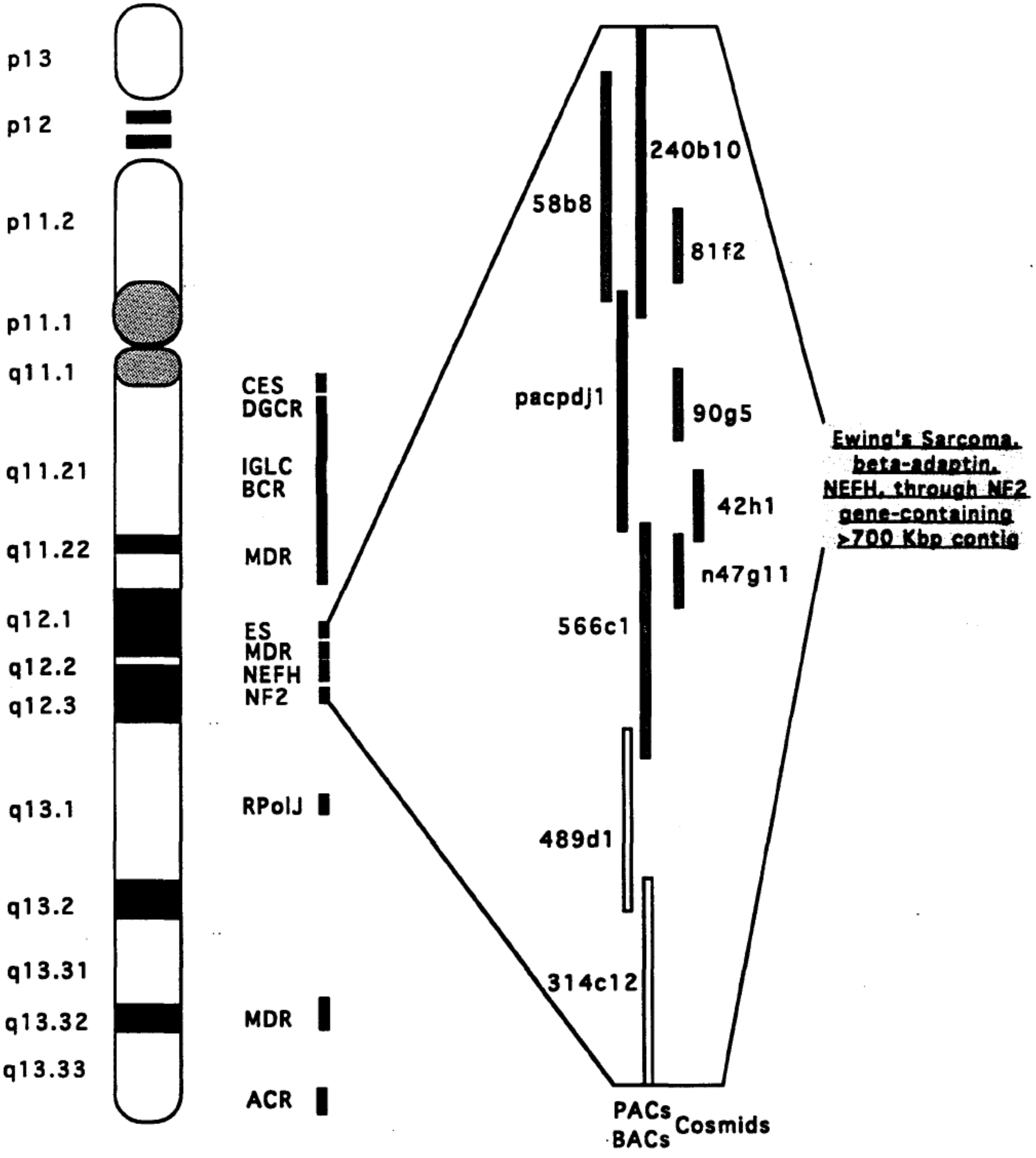
Close contigs by Long Ranger gels, primer walking with fluorescent terminators by Taq cycle sequencing, PCR-based gap amplification followed by shotgun shearing and random sequencing, and/or mapping by sequencing (subclone size selected restriction fragments followed by end sequencing)

Human Chromosome 22 and Syntenic Mouse Chromosomal Regions



CES = Cat Eye Syndrome Region
DGCR = DiGeorge Syndrome Critical Region
IGCL = Immuoglobulin Light Chain Region
GNAZ =Guanine Nucleotide Binding Protein
BCR = Breakpoint Cluster Region
MDR = Meningioma Deletion Regions
ES = Ewing's Sarcoma
NEFH = Neurofilament Heavy Subunit
NF2 = Neurofibroblastoma Region 2
RPolJ = RNA Polymerase II subunit J
ACR = Acrosin

Cosmid, BAC, and PAC clones in the Ewing's Sarcoma through NF2 Regions of Human Chromosome 22

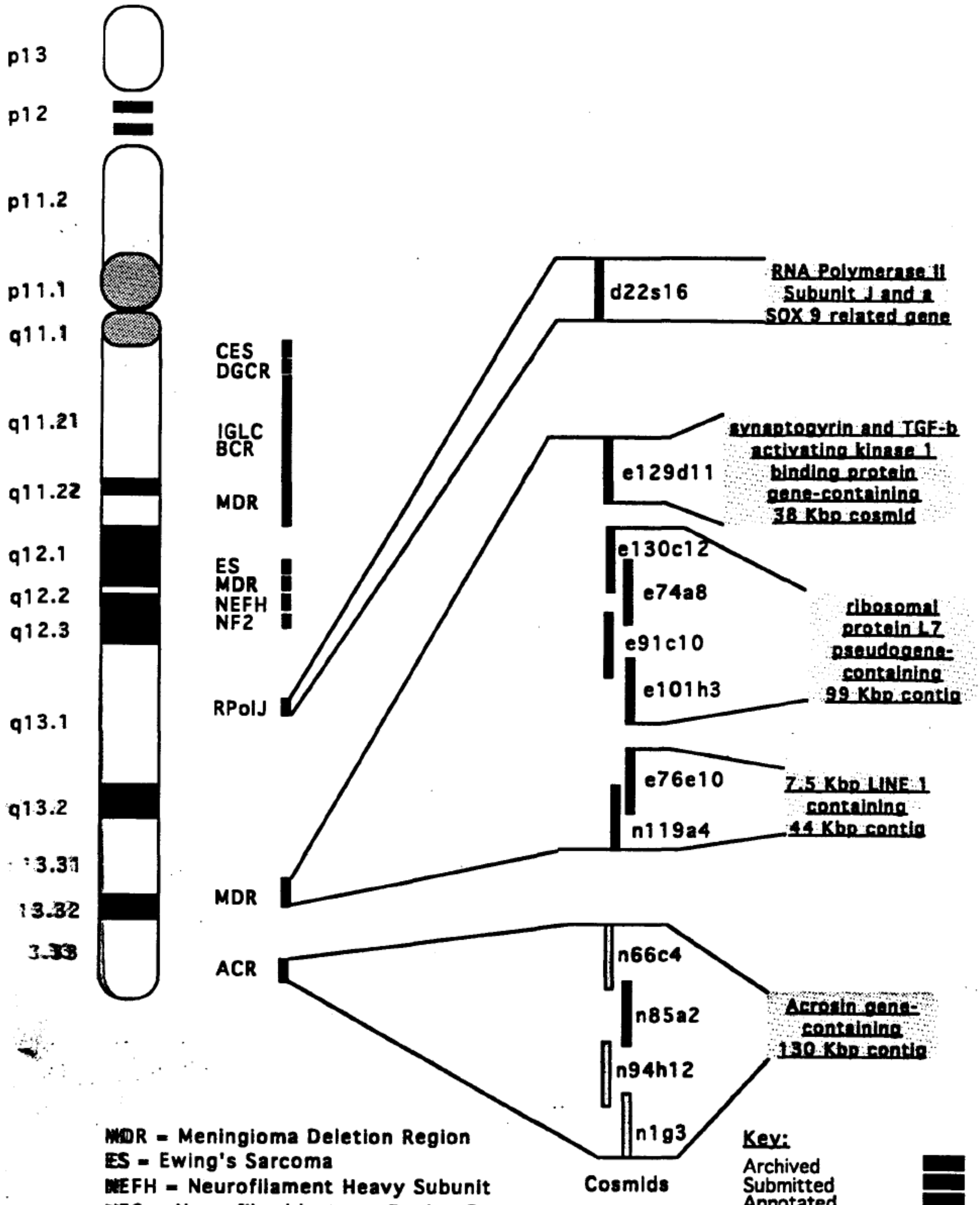


MDR = Meningioma Deletion Region
ES = Ewing's Sarcoma
NEFH = Neurofilament Heavy Subunit
NF2 = Neurofibroblastoma Region 2
RPolJ = RNA Polymerase II subunit J
MDR = Meningioma Deletion Region
ACR = Acrosin

Key:

Archived	■
Submitted	■
Annotated	■
Finished	■
Closure in progress	▨
Shotgun complete	■
Shotgun in progress	■
DNA made	■
Bacterial Clone	□

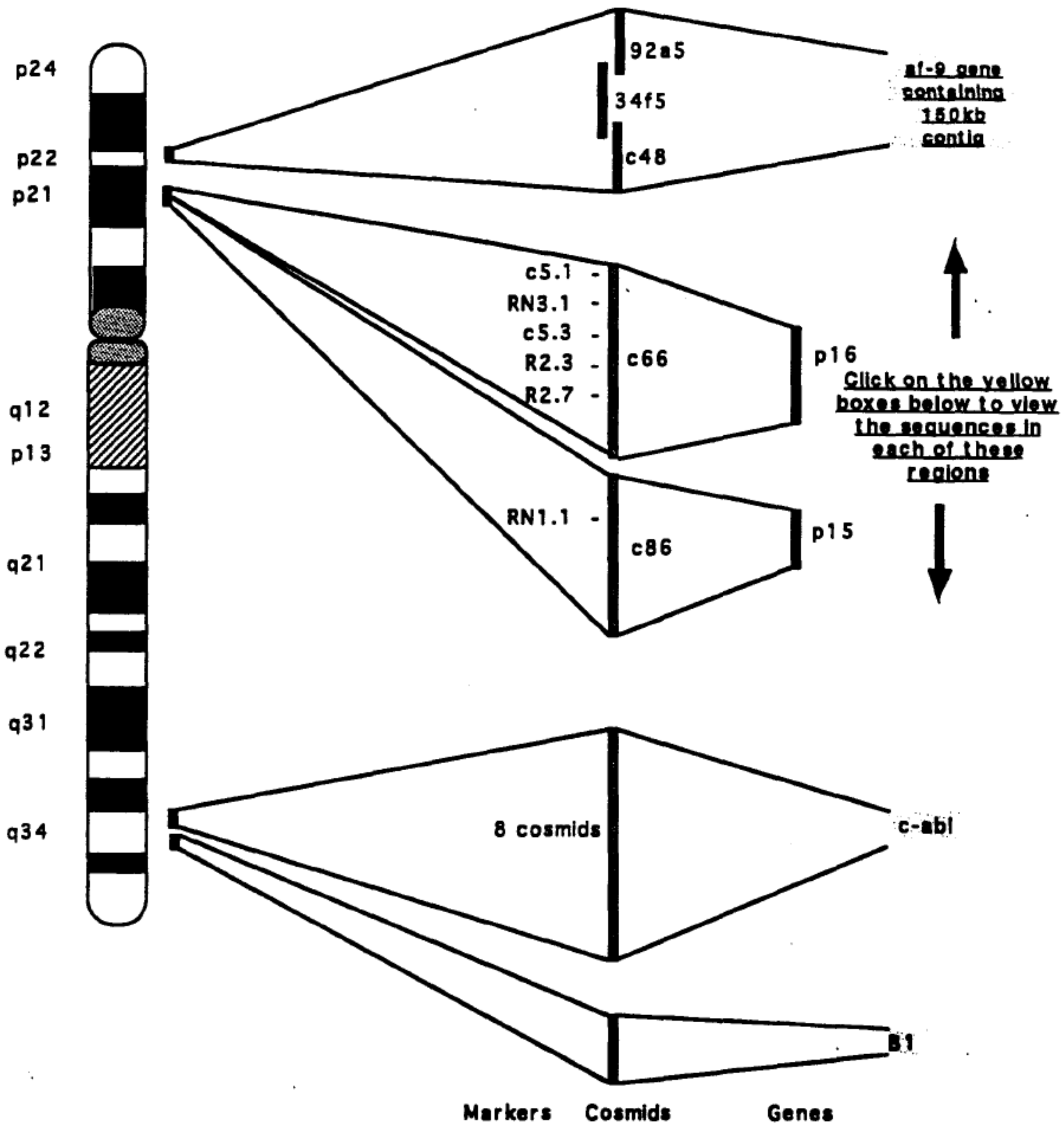
Regions Sequenced at the University of Oklahoma from Clones that Map to the Lower Half of Human Chromosome 22


p13
p12
p11.2
p11.1
q11.1
q11.21
q11.22
q12.1
q12.2
q12.3
q13.1
q13.2
13.31
13.32
3.33
CES
DGCR
IGLC
BCR
MDR
ES
MDR
NEFH
NF2
RPolJ
MDR
ACR
d22s16
RNA Polymerase II Subunit J and a SOX 9 related gene
e129d11
synaptogyrin and TGF-b activating kinase 1 binding protein gene-containing 38 Kbp cosmid
e130c12
e74a8
e91c10
e101h3
ribosomal protein L7 pseudogene-containing 99 Kbp contig
e76e10
n119a4
7.5 Kbp LINE 1 containing 44 Kbp contig
n66c4
n85a2
n94h12
n1g3
Acrosin gene-containing 130 Kbp contig
Cosmids

MDR = Meningioma Deletion Region
ES = Ewing's Sarcoma
NEFH = Neurofilament Heavy Subunit
NF2 = Neurofibroblastoma Region 2
RPolJ = RNA Polymerase II subunit J
MDR = Meningioma Deletion Region
ACR = Acrosin

Key:
Archived ■
Submitted ■
Annotated ■
Finished ■
Closure in progress □
Shotgun complete ■
Shotgun in progress ■
DNA made ■
Bacterial Clone □

Cosmids, and P1's Implicated in Leukemia, Melanoma, and Other Cancers from Human Chromosome 9



Click on the yellow boxes below to view the sequences in each of these regions

Notes:

C48 encodes the portion of the af-9 gene involved in leukemogenic t(9:11) translocations. At least six breakpoints have been mapped to C48.

C66 and C86 encode all.of p16 (CDK-INK4) and p15 (CDK-INK4b), respectively, which, when deleted, are involved in melanomas and other cancers.

Key:

Archived	■
Submitted	■
Annotated	■
Finished	■
Closure in progress	▨
Shotgun complete	■
Shotgun in progress	■
DNA made	■
Bacterial Clone	□

Bacterial Genomes and A. nidulans EST Sequencing Projects

- The initial shotgun sequencing phase of the *Neisseria gonorrohoeae* 2.2 Mbp and *Streptococcus pyogenes 1.9 Mbp* genomes is complete and in closure.

- 95% of each genome is now publicly available on our website. http://www.genome.ou.edu

- *Aspergillus nidulans* EST project is underway.

- Indicates Data Not Available
Level 0 = In Shotgun Level 1 = Unordered Contigs Level 2 = Ordered Contigs
Level 3 = Completely Finished (3-x coverage and fewer than 1 ambiguity/10,000 bases)

Notes Regarding Sequencing Progress:

Maps showing the location of the clones sequenced or in progress are available along with our protocols on our web site:
http://www.genome.ou.edu

All the clones with GenBank accession numbers AC000067 through AC000095 have no gaps and a sequence ambiguity of approximately 5/10,000 bases due mainly to the lack of "rule of three" coverage. These regions presently are being finished by a combination of long gel reads and sequencing off pcr-generated templates prior to declaring that they are at level 3.

It should be noted that to date we have generated:

Total sequence data submitted for 8-31-95 - 9-1-96:	1,846,870 bp	
Total submitted 9-2-96 - 11-15-96:	1,997,137 bp	
Total additional in progress:	2,135,627 bp	*
Total:	5,979,634 bp	*

* = changed since November 15th submission

Conclusions

- The sequence of over 5 Mb of human genomic DNA reveals a gene density of slightly more than 2 genes/100kb, numerious EST's and STS's, and an extensive repeated sequences.
- Although full length alu sequences account for slightly less than 10% of the repeated sequences, the high levels of regions with partial alu sequence homology brings the alu sequence content in these regions to approximately 30-40% with several cosmid sized regions approaching 80% alu homology.
- Continued efforts are in progress to improve the efficiency of large scale genomic level sequencing, data analysis, and data release.

AGENDA
Ninth International Strategy Meeting on Human Genome Sequencing
Cold Spring Harbor Laboratory
Cold Spring Harbor, NY
May 8-9, 2001
DRAFT

May 8, 2001

8:00 p.m. Reception in Blackford Pub

May 9, 2001

8:00 a.m. Continental breakfast in the Plimpton Room, Beckman Laboratory

INTRODUCTORY REMARKS

Morning Session Chair: Francis Collins

DISCUSSION OF HUMAN GENOME DRAFT ASSEMBLIES

Chair: Bob Waterston

8:30 a.m.	Map status	John McPherson
	Current sequence assemblies & plans for updates	Jim Kent Ewan Birney David Lipman
	Plans for updating IGI and IPI	Ewan Birney

10:00 a.m. Coffee Break

FINISHING THE SEQUENCE OF THE HUMAN GENOME

~~**Chair: Francis Collins**~~

10:15 a.m.	Overview of Finishing & report of the finishing working group	Rick Wilson
	Summary of finishing progress & projections	Adam Felsenfeld
	Quality assessment of finished sequence	
	Coordination of individual publications	

12:30 p.m. Lunch

Afternoon Session Chair: Ari Patrinos

PLANS FOR SEQUENCING OTHER LARGE GENOMES

2:00 p.m.	Mouse	Bob Waterston Eric Lander
	Rat	Richard Gibbs
	Zebrafish	Jane Rogers
	Pig	Henry Yang
	Chimpanzee	Masahira Hattori
	Tetraodon	Jean Weissenbach

3:00 p.m. *Coffee Break*

DATA RELEASE ISSUES

3:15 p.m. Discussion of updated data release policy for genome sequencing projects — Eric Lander

NEXT INTERNATIONAL MEETING IN BEIJING

3:45 p.m. Plans for August meeting in Beijing — Henry Yang

SESSION V: INTERNATIONAL SEQUENCING FORUM

4:00 p.m. Summary of Marco Island subcommittee meeting and discussion of plans to convene forum — Francis Collins

4:30 p.m. SUMMARY AND CONCLUSIONS

Ninth International Strategy Meeting on Human Genome Sequencing
Cold Spring Harbor Laboratoy, Cold Spring Harbor, NY, May 9, 2001

Human Sequencing Production Summary

Center	Chrom. Allocation (Mb)	As of April 1, 2001								Clones to top up	Clones to Finish	Additional clones for gap closure
		Total sequence in GenBank (Mb)				Total sequence according to center records (Mb)						
		DRAFT	FINISHED	FULLTOP	ACTIVEFIN	DRAFT	FINISHED	FULLTOP	ACTIVEFIN			
Baylor	323	301	75	52	0							
Beijing	30	22	34	0	0	22.6	37.8	0	1	0	1	unknown
GBF	5	5	2	0	0							
Genoscope	109	76	21	0	0	0	45.9	0	49.3	1	296	
GTC	40	60	10	0	0	29.2	10	8.1	11	165	271	0
IMB	12	31	21	0	0	13	21	12	6	30	80	25
Ins. for Sys. Bio.	32	18	18	1	0	4	17	10	3	30	103	6
JGI/SHGC/LANL	350	251	122	0	0	43	124	62	149	307	750	630
Keio	30	6	15	0	0	7.2	16.4	0	0	0	15	82
MPIMG	5	5	4	0	0							
RIKEN	105	202	31	0	0	202	31	updating	2			
Sanger Centre	900	696	429	63	189	320	471	82	208	1461	2986	300
Stanford		24	5	12	7							
U. Wash	130	3	20	0	0	1.1	22	0.3	0	1076	121	22
Wash U	605	564	223	50	40							
WIBR	465	1,261	51	0	0							
Total	**3,141**	**3,524**	**1,080**	**177**	**236**	**642**	**796**	**174**	**429**	**3,070**	**4,623**	**1,065**

Ninth International Strategy Meeting on Human Genome Sequencing
Cold Spring Harbor Laboratoy, Cold Spring Harbor, NY, May 9, 2001

Expected Finishing (Mb) per Month

								JGI										**Cumulative Total**
MONTH	**BCM**	**Beijing**	**GBF**	**GS**	**GTC**	**IMB**	**ISB**	**SHGC**	**LANL**	**Keio**	**MPIMG**	**RIKEN**	**SC**	**Stanford**	**U. Wash**	**Wash U.**	**WIBR**	
April-2001		0.186		20	2	0.35	2	8	2	1		4.5	30		1			71.0
May-2001		?		20	2	0.35	2	8	3	1		6	30		2			145.4
June-2001		0		10	2	0.35	2	8	4	1		10	30		3.6			216.3
July-2001		0		1	2	0.35	2	8	5	1		10	35		4.5			285.2
August-2001		0		0	2	0.35	2	8	6	1		10	35		4.8			354.3
September-2001		0		0	2	0.35	1	8	7	1		10	35		6.4			425.1
October-2001		0		0	2	0.35	1	8	7	1		10	40		8			502.4
November-2001		0		0	2	0.35	1	8	7	1		10	40		9.6			581.4
December-2001		0		0	2	0.35	0	8	7	1		10	40		9.6			659.3
January-2002		0		0	2	0.35	1	8	7	1		10	40		12			740.7
February-2002		0		0	2	0.35	1	8	7	1		15	40		12			827.0
March-2002		0		0	2	0.35	1	8	7	1		15	40		12			913.4
April-2002		0		0	2	0.35	0	8	7	1		15	30		12			988.7
May-2002		0		0	2	0.35	0	8	7	1		15	20		12			1054.1
June-2002		0		0	2	0.35	0	8	7	1		15	20		8			1115.4
July-2002		0		0	0	0.35	0	8	7	1		15	10		4.8			1161.6
August-2002		0		0	0	0.35	0	8	7	1		15	0		4.8			1197.7
September-2002		0		0	0	0.35	0	8	7	1		15	0		4			1233.1
October-2002		0		0	0	0.35	0	8	7	1		15	0		4			1268.4
November-2002		0		0	0	0.35	0	8	7	1		15	0		3.2			1303.0
December-2002		0		0	0	0.35	0	8	7	1		15	0		3.2			1337.5
Total	**0**	**0.186**	**0**	**51**	**30**	**7.35**	**16**	**300**		**21**	**0**	**255.5**	**515**	**0**	**141.5**	**0**	**0**	**1337.5**

kb	Draft	"Finished"	Chr 21 Adj	Finished	Fulltop	Activefin	**Mb**	Draft	"Finished"	Chr 21 Adj	Finished	Fulltop	Activefin
BCM	300,690	74,872		74,872	52,216	0	BCM	301	75	0	75	52	0
Beijing	22,117	33,924		33,924	0	0	Beijing	22	34	0	34	0	0
GBF	5,320	68	2,036	2,104	0	0	GBF	5	0	2	2	0	0
GS	75,517	21,469		21,469	0	0	GS	76	21	0	21	0	0
GTC	60,364	10,226		10,226	0	0	GTC	60	10	0	10	0	0
IMB	30,688	14,109	6,625	20,734	0	0	IMB	31	14	7	21	0	0
JGI	251,125	121,968		121,968	0	0	JGI	251	122	0	122	0	0
Keio	5,797	8,172	6,827	14,999	0	0	Keio	6	8	7	15	0	0
MPIMG	4,632	2,844	949	3,793	0	0	MPIMG	5	3	1	4	0	0
RIKEN	201,969	14,692	16,646	31,338	0	0	RIKEN	202	15	17	31	0	0
SC	696,240	428,527		428,527	62,645	188,551	SC	696	429	0	429	63	189
SDSTDC	23,949	4,704		4,704	12,101	7,345	SDSTDC	24	5	0	5	12	7
UWGC	2,559	20,141		20,141	318	0	UWGC	3	20	0	20	0	0
UWMSC	17,989	18,175		18,175	527	375	UWMSC	18	18	0	18	1	0
WIBR	1,261,101	50,774		50,774	0	0	WIBR	1,261	51	0	51	0	0
WUGSC	564,236	222,727		222,727	49,548	39,983	WUGSC	564	223	0	223	50	40

Ninth International Strategy Meeting on Human Genome Sequencing
Cold Spring Harbor Laboratoy, Cold Spring Harbor, NY, May 9, 2001

Human Sequencing Production Summary

Center	Chrom. Allocation (Mb)	As of April 1, 2001								Clones to top up	Clones to Finish	Additional clones for gap closure	
		Total sequence in GenBank (Mb)				Total sequence according to center records (Mb)							
		DRAFT	FINISHED	FULLTOP	ACTIVEFIN	DRAFT	FINISHED	FULLTOP	ACTIVEFIN				
Baylor	323	301	75	52	0								No report
Beijing	30	22	34	0	0	22.6	37.8	0	1	0	1	unknown	Good agreement -- what is the status of conversion of draft --> finished; there appears to be no activity
GBF	5	5	2	0	0								No report
Genoscope	109	76	21	0	0	0	45.9	0	49.3	1	296		poor agreement. 76Mb still listed as draft should be either finished (24.9Mb) or active fin (49.3 Mb). No use of new keywords
GTC	40	60	10	0	0	29.2	10	8.1	11	165	271	0	poor agreement. 60 Mb still listed as draft. 29 remain as draft, will that be topped up by summer? Should be 8.1 Mb of full top, 11 Mb of active fin, and 12 Mb unaccounted for. No use of new keywords
IMB	12	31	21	0	0	13	21	12	6	30	80	25	poor agreement. 13 Mb still listed as draft, will that be topped up by summer? Should be 12 Mb of full top and 6 Mb of active fin. No use of new keywords.
Ins. for Sys. Bio.	32	18	18	1	0	4	17	10	3	30	103	6	moderate agreement. Only 4 Mb of draft remain, will that be topped up by summer?10 Mb of full top and 3 Mb of active fin. Little use of new keywords
JGI/SHGC/LANL	350	251	122	0	0	43	124	62	149	307	750	630	Poor agreement. 43 Mb still listed as draft, will that be topped up by summer? 62Mb of full top and 149 Mb of active fin. No use of new keywords.
Keio	30	6	15	0	0	7.2	16.4	0	0	0	15	82	Good agreement -- what is the status of conversion of draft --> finished; there appears to be no activity
MPIMG	5	5	4	0	0								No report
RIKEN	105	202	31	0	0	202	31	updating	2				Good agreement -- what is the status of conversion of draft --> finished; there appears to be no activity
Sanger Centre	900	696	429	63	189	320	471	82	208	1461	2986	300	Good agreement. New keywords being used, although there is a slight lag between center records and database entries. 320 Mb of draft remaining, will that be topped off by summer?
Stanford		24	5	12	7								No report
U. Wash	130	3	20	0	0	1.1	22	0.3	0	1076	121	22	reasonable agreement. New keywords being used. 1.1 Mb of draft remaining, will that be topped off by summer?
Wash U	605	564	223	50	40								No report
WIBR	465	1,261	51	0	0								No report
Total	**3,141**	**3,524**	**1,080**	**177**	**236**	**642**	**796**	**174**	**429**	**3,070**	**4,623**	**1,065**	

Ninth International Strategy Meeting on Human Genome Sequencing
Cold Spring Harbor Laboratoy, Cold Spring Harbor, NY, May 9, 2001

Human Sequencing Production Summary

Center	Chrom. Allocation (Mb)	As of April 1, 2001									Clones to top up	Clones to finish	Additional clones for gap closure
		Total sequence in GenBank (Mb)					Total sequence according to center records (Mb)						
		DRAFT	FINISHED*	FINISHED**	FULLTOP	ACTIVEFIN	DRAFT	FINISHED	FULLTOP	ACTIVEFIN			
Baylor	323	301	75	75	52	0	323	75	50	16	800	1100	160
Beijing	30	22	34	34	0	0	22.6	37.8	0	1	0	1	unknown
GBF	5	5	2	0	0	0							
Genoscope	109	76	21	21	0	0	0	45.9	0	49.3	1	296	
GTC	40	60	10	10	0	0	29.2	10	8.1	11	165	271	0
IMB	12	31	21	14	0	0	13	21	12	6	30	80	25
Ins. for Sys. Bio.	32	18	18	18	1	0	4	17	10	3	30	103	6
JGI/SHGC/LANL	350	251	122	122	0	0	43	124	62	149	307	750	630
Keio	30	6	15	8	0	0	7.2	16.4	0	0	0	15	82
Max-Planck	5	5	4	3	0	0	3.2	4	0	0	0	21	3
RIKEN	105	202	31	15	0	0	77	31	110	15	113	219	40
Sanger Centre	900	696	429	429	63	189	320	471	82	208	1461	2986	300
Stanford		24	5	5	12	7							
Univ. Wash.	130	3	20	20	0	0	1.1	22	0.3	0	1076	121	22
Wash. Univ.	605	564	223	223	50	40							
Whitehead	465	1,261	51	51	0	0							
Total	**3,141**	**3,524**	**1,080**	**1,047**	**177**	**236**	**843**	**875**	**334**	**458**	**3,983**	**5,963**	**1,268**

*credited for chromosome 21
**not credited for chromosome 21

Ninth International Strategy Meeting on Human Genome Sequencing
Cold Spring Harbor Laboratoy, Cold Spring Harbor, NY, May 9, 2001

Expected Finishing (Mb) per Month

								JGI										Cumulative Total
MONTH	BCM	Beijing	GBF	GS	GTC	IMB	ISB	SHGC	LANL	Keio	MPIMG	RIKEN	SC	Stanford	U. Wash	Wash U.	WIBR	
April-2001	11	0.186		20	2	0.35	2	8	2	1	1.1	4.5	30		1	21	10	114.1
May-2001	12	?		20	2	0.35	2	8	3	1	0.2	6	30		2	22	10	232.7
June-2001	13			10	2	0.35	2	8	4	1	0.3	10	30		3.6	23	15	354.9
July-2001	14			1	2	0.35	2	8	5	1	0.3	10	35		4.5	24	20	482.1
August-2001	14				2	0.35	2	8	6	1	0.27	10	35		4.8	25	20	610.5
September-2001	14				2	0.35	1	8	7	1	0.27	10	35		6.4	26	25	746.5
October-2001	14				2	0.35	1	8	7	1	0.27	10	40		8	27	30	895.1
November-2001	14				2	0.35	1	8	7	1	0.27	10	40		9.6	28	35	1051.4
December-2001	14				2	0.35	0	8	7	1	0.27	10	40		9.6	29	40	1212.6
January-2002	14				2	0.35	1	8	7	1	0.27	10	40		12	30	40	1378.2
February-2002	14				2	0.35	1	8	7	1	0.27	15	40		12	31	40	1549.8
March-2002	14				2	0.35	1	8	7	1	0.27	15	40		12	32	40	1722.4
April-2002	14				2	0.35		8	7	1		15	30		12	33	40	1884.8
May-2002	14				2	0.35		8	7	1		15	20		12	34	40	2038.1
June-2002	14				2	0.35		8	7	1		15	20		8	35	40	2188.5
July-2002	14					0.35		8	7	1		15	10		4.8	36	40	2324.6
August-2002	14					0.35		8	7	1		15			4.8	37	40	2451.8
September-2002	14					0.35		8	7	1		15			4	38	40	2579.1
October-2002	14					0.35		8	7	1		15			4	39	40	2707.5
November-2002	14					0.35		8	7	1		15			3.2	40	40	2836.0
December-2002	14					0.35		8	7	1		15			3.2	40	40	2964.6
Total	**288**	**0.186**	**0**	**51**	**30**	**7.35**	**16**	**300**		**21**	**4.06**	**255.5**	**515**	**0**	**141.5**	**650**	**685**	**2964.6**

95% human

Whitehead total finished sequence output

	orig est.	actual	new est.
Dec-00	4	3.2	
Jan-01	4	1.3	
Feb-01	4	4.1	
Mar-01	10	4.2	
Apr-01	10	10.2	
May-01	20		10
Jun-01	20		15
Jul-01	30		20
Aug-01	30		20
Sep-01	40		25
Oct-01	40		30
Nov-01	40		35
Dec-01	40		40

Ninth International Strategy Meeting on Human Genome Sequencing
Cold Spring Harbor Laboratoy, Cold Spring Harbor, NY, May 9, 2001

Human Sequencing Production Summary

Center	Chrom. Allocation (Mb)	As of April 1, 2001								Clones to top up	Clones to Finish	Additional clones for gap closure	
		Total sequence in GenBank (Mb)				Total sequence according to center records (Mb)							
		DRAFT	FINISHED	FULLTOP	ACTIVEFIN	DRAFT	FINISHED	FULLTOP	ACTIVEFIN				
Baylor	323	301	75	52	0								No report
Beijing	30	22	34	0	0	22.6	37.8	0	1	0	1	unknown	Good agreement -- what is the status of conversion of draft --> finished; there appears to be no activity
GBF	5	5	2	0	0								No report
Genoscope	109	76	21	0	0	0	45.9	0	49.3	1	296		poor agreement. 76Mb still listed as draft should be either finished (24.9Mb) or active fin (49.3 Mb). No use of new keywords
GTC	40	60	10	0	0	29.2	10	8.1	11	165	271	0	poor agreement. 60 Mb still listed as draft. 29 remain as draft, will that be topped up by summer? Should be 8.1 Mb of full top, 11 Mb of active fin, and 12 Mb unaccounted for. No use of new keywords
IMB	12	31	21	0	0	13	21	12	6	30	80	25	poor agreement. 13 Mb still listed as draft, will that be topped up by summer? Should be 12 Mb of full top and 6 Mb of active fin. No use of new keywords.
Ins. for Sys. Bio.	32	18	18	1	0	4	17	10	3	30	103	6	moderate agreement. Only 4 Mb of draft remain, will that be topped up by summer?10 Mb of full top and 3 Mb of active fin. Little use of new keywords
JGI/SHGC/LANL	350	251	122	0	0	43	124	62	149	307	750	630	Poor agreement. 43 Mb still listed as draft, will that be topped up by summer? 62Mb of full top and 149 Mb of active fin. No use of new keywords.
Keio	30	6	15	0	0	7.2	16.4	0	0	0	15	82	Good agreement -- what is the status of conversion of draft --> finished; there appears to be no activity
MPIMG	5	5	4	0	0								No report
RIKEN	105	202	31	0	0	202	31	updating	2				Good agreement -- what is the status of conversion of draft --> finished; there appears to be no activity
Sanger Centre	900	696	429	63	189	320	471	82	208	1461	2986	300	Good agreement. New keywords being used, although there is a slight lag between center records and database entries. 320 Mb of draft remaining, will that be topped off by summer?
Stanford		24	5	12	7								No report
U. Wash	130	3	20	0	0	1.1	22	0.3	0	1076	121	22	reasonable agreement. New keywords being used. 1.1 Mb of draft remaining, will that be topped off by summer?
Wash U	605	564	223	50	40								No report
WIBR	465	1,261	51	0	0								No report
Total	**3,141**	**3,524**	**1,080**	**177**	**236**	**642**	**796**	**174**	**429**	**3,070**	**4,623**	**1,065**	

Guyer, Mark (NHGRI)

From: Wetterstrand, Kris (NHGRI)
Sent: Monday, May 07, 2001 9:17 AM
To: 'Bloecker, Helmut'
Cc: Guyer, Mark (NHGRI); Peterson, Jane (NHGRI); Felsenfeld, Adam (NHGRI)
Subject: RE: GBF production report for the Ninth International Sequencing meet ing

Dear Helmut,
The amount of 5 Mb for GBF's allocated territory was taken from previous International meeting data reports. It should include all the territory that your center has completed or will complete, including chromosome 21. This number should be updated, if appropriate, so I will adjust your allocation to 11 Mb. I also take this into account for calculating the amount of finished sequence that your center has deposited in GenBank. I have added your sequence contribution (about 2 Mb, I believe) from chromosome 21 back into the finished category for your center. Please let me know if this is in error. Feel free to contact me with any more questions.
Kris

-----Original Message-----
From: Helmut Blöcker [[redacted]
Sent: Monday, May 07, 2001 4:03 AM
To: Guyer, Mark (NHGRI)
Subject: Re: GBF production report for the Ninth International Sequencing meet ing
Importance: High

Dear Mark,

We have in fact not sent anything yet, but I am about to send the data. Just one question so that I don't mix up things. Up to now we have contributed to chromosomes 21 and 9. In the table I received from you I found the following data

allocated 5Mb
drafted 5Mb
finished 2Mb
fulltopped 0Mb
"activefinned" 0Mb

Does this include, according to your dbsearch, data from chr 9 **AND** chr 21 or do you consider only unfinished chromosomes??

If everything is considered then the allocation should be about 11 Mb. Could you please give me a hint? I would hate to confuse anybody.

See you tomorrow.

Helmut.

****

Dr. Helmut Bloecker
Head of Department
Genome Analysis
GBF (German Research Centre for Biotechnology)

Mascheroder Weg 1
D-38124 Braunschweig
Germany

[redacted]

AGENDA
Ninth International Strategy Meeting on Human Genome Sequencing
Cold Spring Harbor Laboratory
Cold Spring Harbor, NY
May 8-9, 2001

May 8, 2001

8:00 p.m.	Reception in Blackford Pub

May 9, 2001

8:00 a.m.	*Continental breakfast in the Plimpton Room, Beckman Laboratory*

INTRODUCTIONS

DISCUSSION OF HUMAN GENOME DRAFT ASSEMBLIES
Chair: Bob Waterston

8:30 a.m.	Map status	John McPherson
	Current sequence assemblies & plans for updates	Jim Kent Ewan Birney Greg Schuler
	Plans for updating IGI and IPI	Ewan Birney Greg Schuler
10:00 a.m.	*Coffee Break*	

FINISHING THE SEQUENCE OF THE HUMAN GENOME
Chair: Francis Collins

10:15 a.m.	Human Sequence Production	Adam Felsenfeld
	Overview of Finishing & report of the finishing working group	Rick Wilson
	Quality assessment of finished sequence	Adam Felsenfeld
	Coordination of individual publications	
12:30 p.m.	*Lunch*	

Afternoon Session Chair: Ari Patrinos

PLANS FOR SEQUENCING OTHER LARGE GENOMES

1:15 p.m.	Mouse	Bob Waterston Eric Lander
	Rat	Richard Gibbs
	Zebrafish	Jane Rogers
	Pig	Henry Yang
	Chimpanzee	Masahira Hattori
	Tetraodon	Jean Weissenbach

2:15 p.m. *Coffee Break*

DATA RELEASE ISSUES

2:30 p.m. Discussion of updated data release policy for genome sequencing projects — Eric Lander

NEXT INTERNATIONAL MEETING IN BEIJING

3:45 p.m. Plans for August meeting in Hangzhou — Henry Yang

SESSION V: INTERNATIONAL SEQUENCING FORUM

4:00 p.m. Summary of Marco Island subcommittee meeting and discussion of plans to convene forum — Francis Collins

4:30 p.m. SUMMARY AND CONCLUSIONS

Pertinent Points for Scientists and Journal Editors on the Pre-Publication Release of Large Biological Datasets

The Human Genome Project has been notable not just for the efforts by participating scientists to produce large quantities of biological data, but for their decision to make the data available as rapidly and as freely as possible. Early on, several of the genome centers pioneered a new approach by sharing with the community results from large biological projects, while their efforts were still underway and long before the results were published in the scientific literature. The International Human Genome Sequencing Consortium codified this approach in the Bermuda Principles, by which the participants agreed to release sequences assemblies (of 1 - 2 kb and larger) within 24 hours of assembly.

Such pre-publication data release was intended to accelerate scientific progress, by allowing thousands of other investigators to use the findings in their own scientific research without delay. The approach has been remarkably successful, as evidenced by scores of scientific papers that made use of the draft human genome sequence before its formal publication. The benefits to the scientific community of this practice have been widely recognized and it has now been adopted or is being considered for adoption by other large-scale projects involving both sequence and other kinds of data.

At the same time, since this practice is not yet a standard one in science, there has understandably been some confusion concerning the nature of pre-publication data releases:

- Scientists who release pre-publication data from large projects ('participating scientists') often intend to retain the right to publish the first scientific report based on large-scale analyses of their overall dataset. This was, for example, the case in the instance of the human genome sequence.

- Some recent cases suggest, however, that this intention may not be universally understood or accepted. For example, there have been instances in which scientists not participating in a project ('non-participating scientists') have prepared papers that do not simply use the data in their own research (e.g., to study a gene, gene family or region of interest), but which constitute the kind of large-scale analysis of the project's unpublished data for which the participating scientists undertook the project in the first place. In at least one instance, such a paper has undermined the project team's ability to publish its own report.

- Journal editors have been unclear about how to deal with such situations.

Given the unarguable value of pre-publication data release from large-scale biological projects and the desirability of encouraging more participants to adopt such policies, it is important to clarify these issues. Accordingly, we wish to promulgate the following principles as a framework for groups wishing to engage in pre-publication data release.

1. Early release is strongly recommended in light of the benefits to science.
2. Pre-publication deposition of large-scale datasets in publicly available databases is not equivalent to publication.
3. Participating scientists are free to set the terms governing pre-publication release of the data from their project, consistent with any policies required by their funders.
4. These terms should be posted in an appropriate and prominent location, such as the project's web site. The terms should clearly state the (limited) objectives for which the participating scientists intend to use the data, leaving all other uses available to non-participating scientists.
5. Journal editors should expect non-participating scientists to conform to the participating scientists' terms. Specifically, non-participating scientists should show that the use of any pre-publication data is within the terms set by the participating scientists – as would be required for unpublished data used by 'personal communication'. In the vast majority of cases, the situation should be clear and it should be sufficient to simply attach the posted terms. If the situation is not completely clear, the journal editor should expect non-participating scientists to provide an appropriate letter from the participating scientists (as with a 'personal communication').

Although the decision about the terms for data release rest solely with the participating scientists and their funders, we encourage participating scientists to adopt the broadest possible terms and to retain only the right to publish initial large-scale analyses in a timely manner. We also encourage the creation of a regular forum among projects engaged in release of pre-publication data at which issues arising may be discussed and policies refined.

As an example, the International Human Genome Sequencing Consortium has updated and codified terms governing release of pre-publication data in the future. These terms are attached here, and may serve as a useful starting point for other similar consortia.

Updated Policy on the Rapid Release of Human Genomic Sequence Data

International Human Genome Sequencing Consortium
May 2001

To achieve the goal of accelerating research throughout the scientific community through rapid pre-publication data release while respecting the right of sequencing centers to publish primary reports of their projects, the IHGSC adopts the following update to its data release policy:

1. Members of the IHGSC confirm their commitment to release all assembled sequence data rapidly as stated in the 1996 and 1997 Bermuda Principles, i.e. assemblies of 1-2 kb will be released within 24 hours of assembly.
2. For projects generating whole genome shotgun sequence reads, the sequence data (including trace data and relevant ancillary information) will be deposited weekly into the EBI/NCBI Trace Repository.

3. The following terms will govern the use of these pre-publication data, and should be prominently displayed on the center's web sites and on any repositories containing the data.

, [WHAT ORGANISMS DOES THE ABOVE REFER TO? ARE WE REFERRING TO ALL BACTERIA, FUNGAL GENOMES? DO WE REALLY WANT TRACES ON EVERYTHING?]

> "As a public service to the biological research community, DNA sequence data produced by the International Human Genome Sequencing Consortium are made available by the sequence producers before assembly, analysis and scientific publication by deposition in the public international sequence database. The Consortium does so with the understanding that these data depositions and releases do not constitute scientific publication. Furthermore, the depositions are done with the recognition that certain of these data are preliminary and may contain errors and possible contamination (e.g., from yeast and *E. coli* host strains). Once deposited, but prior to the publication by the sequence producers of the complete sequence or large region of the relevant genome, the data are available to all as follows:
>
> a. Users may freely download the data, use them to undertake all types of analyses, and publish scientific papers about these analyses – except as described under (b)
> b. The producing laboratories intend to publish the completed sequence of the genome and certain large-scale analyses of the sequence in a timely manner upon the completion of sequence data acquisition. Therefore, the sole exception to the unrestricted use of these unpublished data is that the data may not be used for the initial publication of large-scale assembly or analysis of the genome. In this context, "large-scale assembly" refers to regions approaching the size of a chromosome or larger and "large-scale analysis" refers to analyses such as identification of genomic features such as genes, repeats, GC content, evolutionarily conserved regions, etc. across those regions. The producing laboratories are, however, open to the possibility of collaboration on such assemblies or analyses.
> c. The data may be repackaged in other databases, provided that appropriate acknowledgement is given to the producer(s) and that this notice describing the terms of use is included."

Guyer, Mark (NHGRI)

From:	Helmut Blöcker [redacted]
Sent:	Monday, May 07, 2001 4:03 AM
To:	Guyer, Mark (NHGRI)
Subject:	Re: GBF production report for the Ninth International Sequencing meet ing
Importance:	High

At 23:16 02.05.2001 -0400, you wrote:

> Dear Helmut:
>
> In reviewing the production data reports from the G16 centers, in preparation for next week's meeting at Cold Spring Harbor, it appears as if we have not received the data that we requested from you recently. If you have sent it and we have misplaced, I apologize. In any case, we would like to ask you to either send us the data before next Monday, May 7, or bring the data with you to the International sequencing meeting so that we can update the summary table directly. Thank you very much and we look forward to seeing you next week.

##

Dear Mark,

We have in fact not sent anything yet, but I am about to send the data. Just one question so that I don't mix up things. Up to now we have contributed to chromosomes 21 and 9. In the table I received from you I found the following data

allocated 5Mb

drafted 5Mb

finished 2Mb

fulltopped 0Mb

"activefinned" 0Mb

Does this include, according to your dbsearch, data from chr 9 **AND** chr 21 or do you consider only unfinished chromosomes??

If everything is considered then the allocation should be about 11 Mb. Could you please give me a hint? I would hate to confuse anybody.

See you tomorrow.

Helmut.

****

Dr. Helmut Bloecker

Head of Department
Genome Analysis
GBF (German Research Centre for Biotechnology)
Mascheroder Weg 1
D-38124 Braunschweig
Germany

[redacted]

Sheet 1

Overall Program Budget

Human Sequencing	FY 01	FY 02	FY 03	Renewal Due	Fund by
Lander	$ 57.9	$ 59.3	$ 60.6	3/03	11/03
Waterston	$ 44.8	$ 49.8	$ 50.6	3/03	11/03
Gibbs	$ 19.4	$ 21.8	$ 22.4	3/03	11/03
Olson	$ 7.3	$ 5.1		3/02	1/03
Smith	$ 5.2	$ 3.5		9/01	7/02*
Davis	$ 3.1	$ 2.1		9/01	7/02*
Hood	$ 4.8				
Subtotal	**$ 142.5**	**$ 141.6**	**$ 133.6**		

Mouse Sequencing	FY 01	FY 02	FY 03	Renewal Due	Fund by
Lander	$ 5.5	$ 6.7		1/02	9/30/02*
McPherson	$ 7.2	$ 8.2		1/02	9/30/02*
Green	$ 5.7	$ 6.8		1/02	10/1/02
Kucherlapati	$ 1.7	$ 1.9		1/02	9/30/02*
McCombie	$ 2.0	$ 2.2		1/02	9/30/02*
Roe	$ 3.4	$ 3.8		1/02	9/30/02*
Subtotal	**$ 25.5**	**$ 29.6**			

Rat Sequencing	FY 01	FY 02	FY 03	Renewal Due	Fund by
NHGRI Funds					
Gibbs	$ 6.4	$ 7.2		1/02	9/30/02*
Smith	$ 3.6	$ 4.0		1/02	9/30/02*
Weiss	$ 1.5	$ 1.6		3/02	11/2/02
Nierman	$ 1.5	$ 1.0			
Marra	$ 0.6	$ 0.8	$ 0.3		
NHGRI Subtotal	**$ 13.6**	**$ 14.6**	**$ 0.3**		
NHLBI Funds					
Gibbs	$ 18.3	$ 19.2			
Holt	$ 12.0	$ 7.0			
NHLBI Subtotal	**$ 30.3**	**$ 26.2**			
Rat Total	**$ 43.9**	**$ 40.8**			

NHGRI Total $ 181.6 $ 185.8 $ 133.6

* Cannot be funded until 11/1/02 with FY 03 funds.

Development of Plan for Mouse Sequencing

Funds already allotted	FY 01	FY 02	FY 03
Human genomic grants - finishing (WI&WU)	**$ 17.1***	**$ 29.6**	**$ -**
Human genomic grant (BCM)	$ 19.4	$ 21.8	
Small human grants (SHOD)	$ 20.4	$ 10.7	
Small mouse grants (Kuch/McC/Roe)	$ 7.1	$ 7.9	$ -
Green	$ 5.7	$ 6.8	$ -
Weiss/Nierman/Marra	$ 3.6	$ 3.4	$ 0.3
Rat genomic grants (BCM&GTC)	$ 10.0	$ 11.2	$ -

* from June 1, 2001 to October 31, 2001.

FY 03 Funds

Funds not yet allotted	FY 01	FY 02	FY 03	Model #1		Model #2		Model #3	
				for Mouse	for other uses	for Mouse	for other uses	for Mouse	for other uses
Human & mouse genomic grants - awarded (WI&WU)	$ 25.6*	$ 94.5	$ -						
Human genomic grants - continued (WI&WU)	$ -	$ -	$ 111.2	$ 111.2		$		$	
Human genomic grant (BCM)	$ -	$ -	$ 22.4	$ -	$22.4 for continued rat funding	$		$	
Mouse genomic grants - renewal (WI&WU)	$ -	$ -	$ 14.9	$ 14.9		$		$	
Small human grants (SOD)	$ -	$ -	$ 10.7	$ 10.7		$		$	
Small mouse grants (Kuch/McC/Roe)	$ -	$ -	$ 7.9	$ 2.9	$5 M for sequencing BACs of other organisms	$		$	
Green	$ -	$ -	$ 6.8	$ -	$6.8 for Green	$		$	
Weiss/Nierman/Marra	$ -	$ -	$ 3.1	$ 3.1		$		$	
Rat genomic grants (BCM&GTC)	$ -	$ -	$ 11.2	$ -	$11.2 for continued rat funding	$		$	
Total	$ 25.6	$ 94.5	$ 188.2	**$ 142.8**					

Mouse Plan A: What we can do with our existing funds: Sequence the maximum number of BACs to "full top", then finish

1 BAC = 200 kb x 20 reads/kb	4,000 reads/BAC		Finish $.06/bp
$2.5/read=	$10,000 per BAC		$12,000/BAC

Shotgun BACS			**Finish BACs**		**Cost**	**Total Needed**	**We have**	**Difference**
FY 01 (6/1/01 - 10/31/01)	5 mos = 2,560 BACs	$ 25.6	FY 01 (6/1/01 - 10/31/01)	0	$ -	$ 25.6	$ 25.6	$ -
FY 02 (11/1/01 - 10/31/02)	12 mos = 9,450 BACs	$ 94.5	FY 02 (11/1/01 - 10/31/02)	0	$ -	$ 94.5	$ 94.5	$ -
							Model #1 for FY 03	
FY 03 (11/1/02 - 10/31/03)	12 mos = 7,990 BACs	$ 79.9	FY 03 (11/1/02 - 10/31/03)	Finish 5,224 BACs*	$ 52.0	$ 131.9	**$ 142.8**	$ 10.9
FY 04 Renewal grant			FY 04 (11/1/03 - 10/31/04)	Finish 7,388 BACs*	$ 88.7	$ 93.9	$ 142.8	$ 48.9
FY 05 (11/1/04 - 10/31/05)			FY 05 (11/1/04 - 10/31/05)	Finish 7,388 BACs*	$ 88.7	$ 93.9	$ 142.8	$ 48.9
Total BACs Topped up	**20,000**	**$ 200.0**	**Total BACs finished**	**20,000**	**$ 229.3**	**$ 439.8**	**$ 548.5**	

*Requires expansion of finishing. Current projection is 4500/year.

Mouse Plan B: Eric & Bob's plan: Sequence 1,000 BACs/month to Full Top (20 reads/kb)

1 BAC = 200 kb x 20 reads/kb	4,000 reads/BAC		
1000 BAC =	4,000,000 reads/mo		Finish $.06/bp
$2.5/read=	$10,000,000 per month		$12,000/BAC

Shotgun BACS			**Finish BACs**		**Cost**	**Total Needed**	**We have**	**Difference**
FY 01 (6/1/01 - 10/31/01)	5 mos = 5,000 BACs	$ 50.0	FY 01 (6/1/01 - 10/31/01)	0		$ 50.0	$ 25.6	$ (24.4)
FY 02 (11/1/01 - 10/31/02)	12 mos = 12,000 BACs	$ 120.0	FY 02 (11/1/01 - 10/31/02)	Finish 1,950 BACs	$ 23.4	$ 143.4	$ 94.5	$ (48.9)
							Model #1 for FY 03	
FY 03 (11/1/02 - 10/31/03)	3 mos = 3,000 BACs	$ 30.0	FY 03 (11/1/02 - 10/31/03)	Finish 6,016 BACs*	$ 72.2	$ 102.2	**$ 142.8**	$ 70.6
FY 04 Renewal grant			FY 04 (11/1/03 - 10/31/04)	Finish 6,016 BACs*	$ 72.2	$ 72.2	$ 142.8	$ 70.6
FY 05 (11/1/04 - 10/31/05)			FY 05 (11/1/04 - 10/31/05)	Finish 6,016 BACs*	$ 72.2	$ 72.2	$ 142.8	$ 70.6
Total BACs Topped up	**20,000**	**$ 200.0**	**Total BACs finished**	**20,000**	**$ 240.0**	**$ 440.0**	**$ 548.5**	

*Requires expansion of finishng. Current projection is 4500/year.

Sheet 4

Calculations of available funds

	FY 01 (6/1/01-10/31/01)		FY 02		FY 03	
	mouse	human	mouse	human	mouse	human
Whitehead						
Budget	0	24.112	6.7	59.3	0	60.6
Needed for Human Finishing		**-9.6**		**-14.4**		**0**
Available for Mouse	0	14.512	6.7	44.9		60.6
Washington Univ.						
Budget	0	18.6	8.2	49.8	0	50.6
Needed for Human Finishing		**-7.5**		**-15.1**		**0**
Available for Mouse	0	11.1	8.2	34.7		50.6
Total available for mouse:	**25.612**		**94.5**		**111.2**	

Note: in FY 01, all mouse funds used for MSC.

FINISHING CAPACITY

(3/30/01)

Human Finishing $M
FY01 (June - Oct)

WI	WU
9.6	**7.5**

Human Finishing $M
needed for FY02

WI	WU
14.4	**15.1**

Finishing Capacity for Mouse FY02

WHITEHEAD			WASHU		
Mb	BACs	$M	Mb	BACs	$M
240	1,200	**14.4**	150	750	**9.0**

Finishing Capacity for Mouse FY03

WHITEHEAD			WASHU		
Mb	BACs	$M	Mb	BACs	$M
480	2,400	**28.8**	430	2,150	**25.8**

		WHITEHEAD					WASHU					BAYLOR				
		Finished (3/26/01)			Territory (Mb)		Finished (3/26/01)			Territory (Mb)		Finished (3/26/01)			Territory (Mb)	
		50			465		215			605		72			323	
		Megabases		BACs	Cost ($M)		Megabases		BACs	Cost ($M)		Megabases		BACs	Cost ($M)	
		Monthly	Total	Monthly	Monthly	Total	Monthly	Total	Monthly	Monthly	Total	Monthly	Total	Monthly	Monthly	Total
FY01	[illegible]	[illegible]	60		0.6		21	236		1.3		11	83		0.7	
	[illegible]	20	80		1.2		22	258		1.3		12	95		0.7	
	6/30/01	20	100		1.2	1.2	23	281		1.4	1.4	13	108		0.8	0.8
	7/31/01	30	130		1.8	3.0	24	305		1.4	2.8	14	122		0.8	1.6
	8/31/01	30	160		1.8	4.8	25	330		1.5	4.3	14	136		0.8	2.5
	9/30/01	40	200		2.4	7.2	26	356		1.6	5.9	14	150		0.8	3.3
	10/31/01	40	240		2.4	9.6	27	383		1.6	7.5	14	164		0.8	4.1
FY02	11/30/01	40	280		2.4	12.0	28	411		1.7	9.2	14	178		0.8	5.0
	12/31/01	40	320		2.4	14.4	29	440		1.7	10.9	14	192		0.8	5.8
	1/31/02	40	360		2.4	16.8	30	470		1.8	12.7	14	206		0.8	6.7
	2/28/02	40	400		2.4	19.2	31	501		1.9	14.6	14	220		0.8	7.5
	3/31/02	40	440		2.4	21.6	32	533		1.9	16.5	14	234		0.8	8.3
	4/30/02	40	**480**		2.4	24.0	33	566		2.0	18.5	14	248		0.8	9.2
	5/31/02	40	520	200	2.4	26.4	34	600		2.0	20.5	14	262		0.8	10.0
	6/30/02	40	560	200	2.4	28.8	35	**635**		2.1	22.6	14	276		0.8	10.9
	7/31/02	40	600	200	2.4	31.2	36	671	180	2.2	24.8	14	290		0.8	11.7
	8/31/02	40	640	200	2.4	33.6	37	708	185	2.2	27.0	14	304		0.8	12.5
	9/30/02	40	680	200	2.4	36.0	38	746	190	2.3	29.3	14	318		0.8	13.4
	10/31/02	40	720	200	2.4	38.4	39	785	195	2.3	31.6	14	**332**		0.8	14.2
FY03	11/30/02	40	760	200	2.4	40.8	40	825	200	2.4	34.0	14	346	70	0.8	15.1
	12/31/02	40	800	200	2.4	43.2	40	865	200	2.4	36.4	14	360	70	0.8	15.9
	1/31/03	40	840	200	2.4	45.6	35	900	175	2.1	38.5	14	374	70	0.8	16.7
	2/28/03	40	880	200	2.4	48.0	35	935	175	2.1	40.6	14	388	70	0.8	17.6
	3/31/03	40	920	200	2.4	50.4	35	970	175	2.1	42.7	14	402	70	0.8	18.4
	4/30/03	40	960	200	2.4	52.8	35	1,005	175	2.1	44.8	14	416	70	0.8	19.3
	5/31/03	40	1,000	200	2.4	55.2	35	1,040	175	2.1	46.9	14	430	70	0.8	20.1
	6/30/03	40	1,040	200	2.4	57.6	35	1,075	175	2.1	49.0	14	444	70	0.8	20.9
	7/31/03	40	1,080	200	2.4	60.0	35	1,110	175	2.1	51.1	14	458	70	0.8	21.8
	8/31/03	40	1,120	200	2.4	62.4	35	1,145	175	2.1	53.2	14	472	70	0.8	22.6
	9/30/03	40	1,160	200	2.4	64.8	35	1,180	175	2.1	55.3	14	486	70	0.8	23.5
	10/31/03	40	1,200	200	2.4	67.2	35	1,215	175	2.1	57.4	14	500	70	0.8	24.3

$ 0.06 Cost of finishing per base

200 kb per mouse BAC

200 kb per rat BAC

DEPARTMENT OF HEALTH & HUMAN SERVICES

Public Health Service

National Institutes of Health
National Human Genome
Research Institute
31 Center Drive MSC 2033
Building 31, Room B2B07
Bethesda, MD 20892-2033
(301) 496-7531
FAX (301) 480-2770

May 5, 2000

Dear International Meeting Attendee,

In looking over the production of the International Sequencing Consortium over the past quarter, I am struck by the sequencing capacity that has now been built in the public sector laboratories worldwide. I am very much looking forward to our meeting this coming Wednesday, May 10, 2000, as I think it will be an opportunity for us to celebrate this success and plan for the future.

I am enclosing a number of documents as background for the meeting that I hope you will have an opportunity to look over before the meeting:

1. The Agenda
2. The Sequencing Production Summary table
3. A graph showing the accumulation of GenBank deposits for all centers
4. A table of the Working Draft Specifications that were agreed upon at the Sixth International Meeting- we request that you come to the meeting prepared to discuss the status of your database deposits with respect to your implementation of the specifications given in the table. We plan to fill in the table during the discussion at the meeting.
5. Two documents entitled "Proposed Finishing Standards for the Human Genome Project" and "Proposed Finishing Vocabulary". These documents will be discussed during the meeting.
6. A proposed Restatement of HGP Policy for Rapid Data Release of Genomic DNA Sequence

In anticipation of the publication of the working draft sequence later this year, a video documentary of the Human Genome Project is being developed. This video documentary is intended for high school students and the lay public to increase their awareness and understanding of the HGP and its implications for research and medicine. In order to capture the international nature of the Project, we would like to film for about an hour during the May 10th meeting. Thus, when you arrive at the meeting, cameras, lights etc. will be set up and we will ask you to sign a video release form giving your permission to be filmed. Signing the release is completely voluntary and if you are uncomfortable with providing this permission, neither your image nor your words will be included in the video. In spite of the distractions during the meeting, you should feel free to speak freely and try to ignore the crew as much as possible. You need not worry about confidential information being captured and used in the video as it will be extensively edited and reviewed by NHGRI staff to ensure that content of the video is appropriate for public dissemination. We appreciate your willingness in facilitating this exciting video project.

As noted earlier, there will be a reception for attendees in the Blackford Pub starting at 8 PM on the evening of May 9. I hope that you will arrive in time to take advantage of this opportunity to meet with the rest of the attendees before the meeting starts the next morning. The meeting on the 10th will be in the Plimpton Room in the Beckman Laboratory.

I look forward to seeing you at Cold Spring Harbor.

Sincerely,

Francis S Collins

Francis S. Collins, M.D., Ph.D.
Director

Attachments

AGENDA
Seventh International Strategy Meeting on Human Genome Sequencing
Cold Spring Harbor, NY

<u>May 9, 2000</u>

8:00 p.m. Reception in Blackford Pub

<u>May 10, 1000</u>

8:00 a.m. Continental breakfast, Plimpton Room, Beckman Laboratory

8:30 a.m. Welcome: Marv Frazier

SESSION I: UPDATE ON HUMAN GENOME MAP

8:45 a.m. Co-chairs: John McPherson & David Bentley

9:30 a.m. Coffee Break

SESSION II: WORKING DRAFT SEQUENCE

9:45 A.M. Co-chairs: Francis Collins & Marv Frazier

- Review of production since January, 2000 and expectation for late-May or June completion
- Ensuring that the working draft meets all of the previously defined specifications (identified with htgs_draft keyword, quality scores included, end contigs identified, contamination removed, and 100 'N's used to represent gaps in sequence)

SESSION III: WORKING DRAFT PUBLICATION

11:00a.m. Discussion of publication plans and draft
Co-chairs: Francis Collins & John McPherson

12:00 Noon Working lunch

SESSION IV: WHAT HAPPENS AFTER THE WORKING DRAFT IS COMPLETED?

1:00 p.m. Co-chairs: Bob Waterston & Rick Wilson

- Status of chromosomes 21 & 20 (Andre Rosenthal, Yoshi Sakaki & Sanger)
- The next version of the working draft; addition of order and orientation
- Standards and common vocabulary for finished sequence – finishing working group report
- Data release
- Timeline for finishing the Human Genome

4:00 p. m. Summary and Conclusions

Seventh International Strategy Meeting on Human Genome Sequencing
Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, May 10, 2000

Sequencing Production Summary (12/1/99 to 2/29/00)

Center	Regions	Size (Mb)	Yearly (4/1/99 - 3/31/00) Projected Attempted Reads (k)		Last Quarter (12/1/99 - 2/29/00)						Current Quarter (3/1/00 - 5/31/00)		GenBank Totals		
			Jan 2000 Estimates	Updated Estimates	Projected Attempted Reads (k)	Actual Attempted Reads (k)	Actual Successful Reads (k)	Reported Genbank Submissions (kb)	Actual Genbank Submissions (kb)*	% Working Draft	Projected Attempted Reads (k)	Projected GenBank Submissions (kb)	as of 3/3/00 (kb)*	as of 5/1/00 (kb)*	projected for 5/31/00 (kb)***
Baylor	3, 12, X	230	2,405	2,315	870	811	587	67,660	67,660 **	95	1,070	75,000	135,588	176,104	210,588
Beijing	3p	30	500	500	280	280	260	16,618	15,046	99	200	20,000	15,593	26,584	31,593
GBF	9, 21	6	90		25			122	122			1,000	2,663	3,145	3,463
Genoscope	14	85	1,400		400	396	311	26,000	20,088	60	400	25,000	45,280	57,160	65,280
GTC	10	50	450	450	405	328	262	20,700	13,964	72	375	19,000	27,569	41,768	42,769
IMB	8,21,X	50	1,500	720	180	180	175	10,575	6,678	78	180	12,000	26,162	32,581	35,762
JGI	5,16,19	250	5,546	5,776	2,313	2,543	1,823	87,642	87,642 **	92	2,419	89,600	197,158	285,013	286,758
Keio	2,6,8,21,22	23	230	440	110	110	99	5,386	5,414	39	110	6,000	11,498	16,195	16,298
MPIMG	17,21,X	7			20	23	20	2,096	1,904	96	23	2,000	5,573	5,739	7,173
RIKEN	11q, 18, 21	160	2,100	1,800	1,295	939	825	67,817	45,139	90	1,600	75,000	80,351	125,715	140,351
Sanger Centre	1,6,9,10,13, 20,22, X	1,000	10,517	9,645	4,743	4,031	2,674	213,489	213,489 **	93	6,600	360,000	386,271	477,337	746,271
Stanford	8		240	158	137	116	92	9,847	7,004	76	142	11,350	13,588	19,110	22,668
U. Wash (Hood)	14,15	21	170	211	60	79	56	6,918	5,064	82	60	5,500	16,801	21,346	21,201
U. Wash (Olson)	7p, 7q	25	500		125	130	94	5,536	6,848	90	50	5,000	20,386	20,526	24,386
Wash U	2,3,4,7,8,11 ,15,17,18,Y	1,050	6,824	8,663	2,400	2,239	1,587	176,960	176,960 **	97	2,400	240,000	408,581	643,319	648,581
WIBR			10,387	11,051	4,726	5,124	3,980	302,592	302,592 **	99	4,950	357,000	362,372	449,341	719,372
Total			42,859	41,729	18,089	17,328	12,846	1,019,958	975,614		20,579	1,303,450	1,755,434	2,400,982	3,022,514

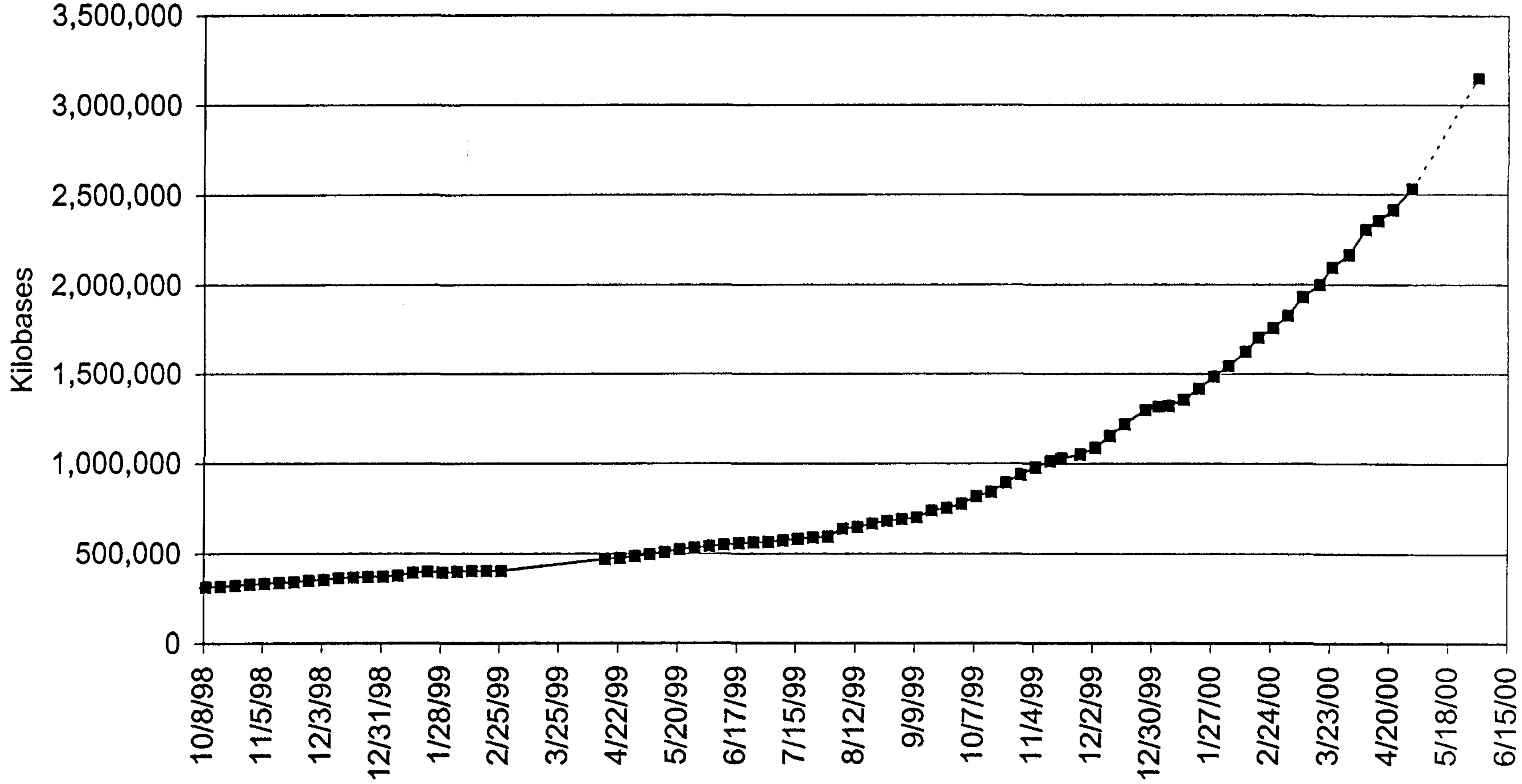
* GenBank totals are taken from combined finished plus draft sequence, where the draft category is reduced by a certain percentage to account for redundant sequence within the database and by 4% due to "N's" in the working draft. On 3/3/00 the level of redundancy was estimated to be 16% and on 5/1/00 the level of redundancy was estimated to be 20%.

** Deposits reported were taken from the GenBank report, therefore reported deposits and actual deposits are the same for these centers.

*** This projection was calculated by adding the base projections from the centers, less 20% to account for working draft redundancy, for the current quarter to the amount of non-redundant sequence in

Seventh International Strategy Meeting on Human Genome Sequencing
Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, May 10, 2000

Total GenBank Deposits*



*Total = "Finished" + "Unfinished" x Reduction %
The unfinished category is reduced by a certain percentage to account for redundant sequence within the database and by 4% due to "N's" in the working draft.

Seventh International Strategy Meeting on Human Genome Sequencing
Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, May 10, 2000

Working Draft Specifications

The status of the following five working draft specifications will be discussed at the International Meeting. Please be prepared to provide information on the status by reporting on the percent of your center's projects meeting the specifications and the your center's timeline for completion on all projects.

Genome Center	"htgs_draft" Keyword Submitted	Quality Scores Included	End Contigs Identified	Non-Human Sequence Removed	100 'N's Used to Represent Gaps in Draft
Baylor					
Beijing					
GBF					
Genoscope					
GTC					
IMB					
JGI					
Keio					
MPIMG					
RIKEN					
Sanger Centre					
Stanford					
U. Wash (Hood)					
U. Wash (Olson)					
Wash U					
WIBR					

Attachment 5: Proposed Finishing Standards and Vocabulary

To begin to discuss the issues that will face the public sequencing effort as it moves into the finishing phase of the human sequencing effort, staff from the Sanger Centre, JGI, and the NHGRI-funded centers have formed a finishing working group. The purpose of the group is to anticipate issues that may arise in increasing finishing production, both within individual centers, and among the several centers contributing to the public effort, and to suggest appropriate ways to address such issues.

As its first main goal, the working group has been discussing two issues. The first is a set of proposed uniform practices for groups to apply to various problem regions (for example: tandem repeats or single-stranded regions) that will be encountered in the finishing of a sequencing project. The second is a standard annotation vocabulary, to ensure that critical terms employed in finishing discussions convey the same meaning to all participants, and to ensure that the various classes of finishing problems that may be encountered receive consistent annotation in the public database. The aim of these proposals is to encourage a presentation of the finished human sequence that is of relatively uniform quality, and ultimately to make it more usable.

The working group's proposals are attached as 'Proposed Finishing Rules' and 'Proposed Finishing Vocabulary'. On those items where the working group has not been able to reach a unanimous view, the main discussion points have been summarized. We appreciate your consideration of all of these proposals, and hope you agree that, even though finishing practices differ between centers, the finished sequence should be of uniform quality across the genome. We look forward to hearing a discussion of these proposals during the meeting.

PROPOSED FINISHING STANDARDS: RECOMMENDATIONS FOR THE HUMAN GENOME PROJECT

STATEMENT TO BE INCLUDED WITH SUBMISSION:

> "This sequence was finished as follows unless otherwise noted: all regions were either double-stranded or sequenced with an alternate chemistry or covered by high quality data (i.e., phred quality >= 30); an attempt was made to resolve all sequencing problems, such as compressions and repeats; all regions were covered by at least one plasmid subclone or more than one M13 subclone; and the assembly was confirmed by restriction digest."

If a sequence meets the criteria of the above statement, it needs no comments or tags. If the criteria are not met, such as ambiguous bases (but we are fairly certain that the sequence is correct), then the region is duly annotated. If we know that the sequence is not resolved unambiguously - such as within a tandem repeat - then an annotation tag is required. [see ANNOTATION VOCABULARY]

It was agreed that the steps listed here are considered to be those undertaken by finishing staff, before the problem is brought to a more experienced finisher or coordinator for approval.

SPECIFIC RULES FOR PROBLEM REGIONS:

In any of the following examples, if there is evidence of unique sequence or genes within the problem region, then every attempt must be made to represent this data in the submitted sequence. The rules apply only in cases where the genome center has made appropriate attempts to resolve the region.

1. large unresolvable tandem repeats (>5Kb)

   Even if the tandem repeat is resolved, the correct representation of the clone is still questionable, since the probability of deletions in large tandems is high.

   Rules for finishing tandem repeats ("annotation tag"):
   a. The repeat must be anchored to unique data on both sides of the region.
   b. Attempt to size the repeat region using restriction digest or PCR.
   c. An attempt must be made to sort all orphan sequence reads.
   d. All other repeat-containing contigs must be checked for unique data.
   e. Force-join anchored contigs without adding Ns.
   f. Clearly annotate the size and nature of the repeat.

2. small unresolvable tandem repeats (<5Kb)

The same rules apply here as for finishing large tandem repeats (annotation tag), except that the region will be sized by PCR or restriction digest.

3. imperfect di/trinucleotide [tandem] repeats
The same rules apply here as for finishing large tandem repeats (annotation tag), except that the region will be sized by PCR or restriction digest.

4. homopolymeric runs
Size the region by restriction digest or PCR (must produce a single, unambiguous product!). If more than 300 bp are missing, then attempts must be made to obtain the missing sequence. If less than 300 bp are missing, and/or if the sequence pattern can be visualized in the traces on both sides of the gap, then force join and annotate in submission.

5. large duplications
These regions can typically be resolved. Stringent assembly parameters and restriction digest data will be helpful. When they cannot be completely resolved, base pair differences between repeat copies must be noted, when they have been detected. Selected reverse primer reads or transposons can be used to confirm which subclones lie within a particular copy of the duplicated sequence.

6. Inverted repeats & "hairpins"
Stringent assembly parameters and selected reverse primer reads should be used to correctly anchor the repeats to unique sequence. Shatter libraries or transposons should be used to provide the sequence of a unique loop.

GENERAL STANDARDS:

1. Extra Data Rule: If the sequence of a finished genomic clone adds at least 100 bases to a previously finished neighbor (i.e., probable deletion in the neighbor), then this data must be submitted to GenBank with the appropriate annotation.

2. A comment is required for regions that are covered by PCR only (including PCR products from subclones). The region should be represented by at least two different PCR products. If only one PCR product is available, the sequence should be annotated to indicate this.

3. Finished sequences must be screened for bacterial transposons. All transposon insertions will be excised from the surrounding human sequence prior to database submission. The insertion site must be annotated and include the size and sequence of the excised transposon.

4. All extra contigs (>1 Kb) within a database must be accounted for.

5. Single-stranded or single-chemistry (i.e., Big Dye terminators) coverage on PCR products is acceptable only if the region passes at greater than phred30 quality. Such coverage is limited to 1% of a genomic clone. All such regions must be annotated.

6. Single-stranded or single-chemistry coverage (whether from subclones or a PCR product) that does not exceed phred30 quality may be passed after formal approval, and must be duly annotated (implies that the genome center has attempted to resolve the region). Such coverage is limited to 1% of a genomic clone.

7. A single plasmid read can pass as single clone and single chemistry, provided that the sequence quality is greater than phred30 throughout the region, and that the region is no larger than 100 bp. The region should be annotated to reflect the single clone coverage. Such coverage is limited to 1% of a genomic clone.

8. Single M13 subclone region is permitted as long a restriction digest or PCR confirms the assembly. The region should be annotated to reflect the single clone coverage. Such coverage is limited to 1% of a genomic clone.

9. A sequence read that provides confirmation in a region of single clone/strand/chemistry coverage need only demonstrate that the primary subclone is not chimeric or deleted. Such coverage is limited to 1% of a genomic clone.

10. "Standard" Taq DNA Polymerase alone should not be used to size or finish unresolved regions, The use of high fidelity DNA polymerases (e.g., "KlenTaq") is required in these regions.

11. When using a restriction digest to size a region other than tandem repeats, the region in question must be contained in a fragment of 3 Kb or less. If this is not possible, multiple digests must be used to confirm the size.

PROPOSED FINISHING VOCABULARY
05/03/00

Below are listed common problems that are encountered in finishing sequence, and suggested annotation vocabulary (in italics) for each.

1. Ambiguous bases: Bases for which we cannot be certain of the consensus. The base should be called as the best guess of the finisher and annotated as *"unsure"*. Ns should not be used in the consensus of finished sequence. Additional comments can be made.

2. Single stranded regions: Regions covered by sequence from one subclone only should be annotated as *"single clone coverage"*. Regions that are covered by sequence from one strand only and with one type of sequencing chemistry need not be annotated if the consensus is covered by a phred 30 base in terminator chemistry at each position.

3a. Overlaps with another large insert clone: Regions of a phase 3 submitted consensus that overlap another phase 3 submitted consensus. Annotation should be precise (*"clone overlap"*) and include as much information about the nature of the overlap as possible. Whenever possible, start, end, position and size of overlapping region should be given. Accession numbers of overlapping clones should be given. It is encouraged to submit at least 200 bases of overlapping sequence.

3b. Partial submission: In the case that a phase 3 submission overlaps another phase 3 submission the annotation should indicate that sequence overlapping with another submission (identified by accession number) was not submitted (*"clone overlap not submitted"*). If a polymorphism that adds a significant amount of data (>100 bp) is known to exist in the unsubmitted overlapping region it should be indicated by annotation.

4. Gaps: Regions of non-contiguity. It is assumed that all reasonable effort to resolve these has been made before submission of gapped sequence is considered.

- **Unresolved tandem repeats (VNTRs):** The plurality advised making the best estimate of size of the repeat region, force joining, and adding an annotation indicating the estimated number of copies of the repeat and any estimate of the number of bases that may not be represented in the submission. The term *"unresolved tandem repeat"* should be used. N's should not be used; they can imply a lower quality of sequence than actually exists.

  Gaps in tandem repeat regions should not be considered to be the same as gaps where no sequence information exists, since in the former one knows the context and unit-repeat sequence, but not the number of repeats.

  In discussion, another view about these repeats held that after an estimate of the repeat region was made, N's should be inserted, rather than making a force join. Annotation would otherwise be the same.

  GC-rich regions should be finished, not left as gaps.

- **Gaps other than those in tandem repeats:** The Sanger Centre has argued strongly that no gapped sequences (other than tandem repeats) be submitted as phase 3. They feel that both representation of sized gaps by an appropriate number of Ns and submission of a gapped project as two separate submissions are not acceptable, and that gapped sequences should remain as phase 2. There is not universal agreement on this point, as some centers represent sized gaps by a string of Ns, while others force join contig ends and annotate. Further discussion is warranted.

- **Remaining for discussion:** Unresolved di- and tri-nucleotide repeats, long mononucleotide runs, low sequence complexity regions [eg. long runs of degenerate, not- quite repetitive GA/CT]. These issues have not yet been discussed in detail by the working group. Considerations are the costs (financial and otherwise) of finishing vs. not finishing these regions. If these regions remain unfinished, but work has ceased on them, it seems like there ought to be a status indicating this fact.

5. PCR template only: Regions of the consensus for which no subclone template was identified, and which was sequenced only from PCR products.

Any region derived from PCR product sequence only should be annotated, and template source given ("*pcr product sequence only*").

6. Bacterial transposon insertion: Bacterial insertion sequences identified in the large insert clone should be excised and the excision point annotated. ("*bacterial transposon excised*") If at all possible the full sequence of the excised region should be given since insertion sequences can be polymorphic, although at present field size limitations preclude simply putting the sequence in a comment. The databases will need to help in deciding how to represent insertion sequences.

7. Other: Other features that might be deemed useful to the user public such as repeats, STSs, regions of similarity, genes, or regions of low quality data may be annotated. Such annotations are optional and done at the discretion of the finisher.

RESTATEMENT OF HGP POLICY FOR
RAPID DATA RELEASE OF GENOMIC DNA SEQUENCE

The public sequencing effort has been adhering to the concept that human genomic DNA sequence produced by large-scale DNA sequencing centers should release data as rapidly as possible. We have publicly stated that

> "in the case of unfinished data...sequence assemblies of 1-2 kb in size should be released within 24 hr of generation."

Now that we are about to complete the working draft, this formulation of the rapid data release policy for "unfinished" sequence is no longer appropriate. Instead, NHGRI proposes that the policy for post-working draft, but still unfinished, sequence should be updated to state that

> *" each time new sequence data are added to a project by rerunning the assembly program, the public database entry should be updated within 24 hours"*

As for finished sequence, we had previously stated that

> "Finished data should be submitted to the public sequence databases on a similarly rapid time scale." [i.e. similar to a 24 hour release time scale]

This was relatively non-specific, but it is not clear to us how to improve it. Even a statement that appears to be clear such as

> *"Projects should be submitted to a public sequence database within 24 hours of being finished."*

begs the question because producers have a considerably more leeway in determining when a project is actually finished than in determining when a sequence assembly was done. We would appreciate advice on a clear policy statement to ensure rapid release of finished sequence data.

HUMAN GENOME SEQUENCING

roup	Cumulative -Feb. '97	Cumulative -Feb '98	ACTUAL '97	CLAIMED '97	PREDICTE '97
anger	14.6 Mb.	36.0	21.4	22	35
T. Louis	4.8	24.5	19.7	20	24
hitehead	2.1	7.0	4.9	7.0	20
IGR	2.7	6.4	3.7	4	11
aylor	3	6.5	3.5	4.6	12
tanford	0.3	0.65	0.35		5
JW	0.59	3.65	3.06	3.2	6
nen-ABI	2.4	5	2.6		3.5
OE	4	4.2	0.2		20
e-Ok.	3.8	4.1	0.3		5.5
ena	1.5	5.29	3.79	3.7	6
TSW	1.6	4.35	2.75		5
Kaki-Jam	2.7	5.1	2.4		3.4

TOTALS

Cumulative -Feb'97	Cumulative -Feb 98	ACTUAL '97	PREDICTED '97
44.1	112.7	$\boxed{68.6}$	158

"Optimism Factor"

$$= \frac{158}{68.6} \approx 2.4$$

Bermuda III

First morning session (Jim Watson, chair)

"Many of the faces here are the same, except for one physicist, who I hope still has the arrogance physicists once had, to point out to us what we are not doing right."

Rick Wilson – collated e-mail responses

How much sequencing has already been done?
GenBank – just over 100 Mb
Amounts reported for this meeting

Sanger	34.85
Wash U	22.02
TIGR	7.0
BCM	6.4
ACGT-ABI	5.0
UTSW	4.35
Washington	3.65
Tokyo	2.9
Stanford	0.65
Oklahoma	4.1
Whitehead	

Stanford

- Longest contigs are single BACs
- Error rate 1/65,000 predicted by phred/phrap analysis
- Ca. 5 Mb unfinished, 3.5 of which is good quality (rest is old which is being set aside, may come back to later); will finish 1.5 Mb of this by April
- Rate of 7.5-8Mb/yr finished
- Clones: de Jong library is adequate so far
- Future plans: starting April 1, want to produce 12 Mb/yr for that year, 22 Mb the following year, and 30 Mb/yr after that; have procured sufficient space
- All sequencing we want to do is on chromosome 4

Critical issues:

- long-range continuity, using high resolution RH hybrid map to build contigs and this is working very well (80 kb resolution); have 2.5 Mb contig built solely w/ RH map and de Jong library, 50 kb STS over this region, 5X depth over the contig
- Second issue is clone validation

DC: very interested in figuring out whether you can get long-range continuity
RH map allows you to tell whether there are gaps in the clone map

Working in a 10 Mb region with half the density of STSs, currently pulling clones
Will allow us to know where the gaps are, we don't have a good idea (nor does anyone else) of how to fill the gaps.

AR: given production of 0.65 Mb, the 20X ramp to 12 Mb next year won't happen
RM: see you in Bermuda next year, actually current rate is 6-8 Mb/yr, not 0.65
DC: is scale from 7 to 12 Mb/yr reasonable?
AR: yes, but I take what there is in GenBank, not words
DC: agree, take on face of what you see, not what you say. But in our case, what present throughput at a certain quality is has to be taken into account, even though it is harder to asses

Level of shotgun?
3.5 in 3kb clones to give initial assembly
reads per kb is better estimate. Now we are below 20, approaching 15 (12-13 in original set, then some finishing)

Univ. of Tokyo

2.9 submitted, 7 Mb actually finished (has been sent to agency and to chromosome 21 Web site, should be in GenBank soon)

at least 1 Mb contigs (minimal accepted by funding agency)
unfinished is about 5 Mb (very fragmented, raw data are in DDBJ)

clone resources: own chromosome 21 P1 library from cell line, others from de Jong PACs, Soeda cosmids

quality: believe is quite high, using nested deletion strategy, 99.9% accurate compared to other chromosome 21 data in GenBank

expect 5 Mb next year, want to finish chromosome 21 by end of century which means another 5 Mb the following year (their part is 1/3 of chromosome). Starting on mouse syntenic regions (mouse chromosome 16)

looking next to chromosome 11

group is moving to new JST sequencing center at Riken next year. Assuming this will mean more money, would like to scale to 8, 15, and finally 30

critical issues: space (will be OK in new center); nested deletion strategy is OK for 10 Mb/yr throughput but not higher, will need new or mixed strategy to go to 30

Univ. of Washington

Average submitted contig is 640 kb, longest is 1.7 Mb
Best estimate is 1/480kb (based on sequencing ca. 240 kb twice in overlapping cosmids, found 4 discrepancies, of which 3 were cosmid mutations)

Also have submitted 400 kb of mouse

Emphasizing continuity, so paying a penalty now in terms of overlap (which other groups are not paying yet)

Haven't found any regions covered by sequence-ready clones that can't be sequenced (contigs will be up to 2.2 Mb when gaps are filled)

Unfinished is under a megabase

Clones: so far have been taking YACs to two-fold coverage, Subcloning into cosmids and making MCD maps (resolution is having an ordered site every 2 kb or so); will be switching to BACs and treating them in the same way as YACs (subcloning into cosmids) as well as shotgunning BACs directly. Will continue to rely on MCD maps, although there are problems with MCD mapping of BACs

There have been some regions that are problematic for subcloning from YACs, have been addressed by subcloning from BACs or PCRing directly from YACs

Currently at 10,000 reads/mo (4000 for Pseudomonas project, will wind down by 6/1 and will convert that capacity to human); 7/1/98 to 7/1/99 hope to produce 10 Mb, and then to reach total of 100 Mb over the next three years

Critical issues: out of space in the university, have to go outside. Setting up company which will be totally dependent on NIH funding (which will mean NIH support for full cost of sequencing). This is also an issue of retaining good people (four people are running the operation and their prospects are not good for obtaining academic positions doing this kind of work).

FC: NIH has no problem with support for private sector, issue is one of relative cost.

Univ. of Oklahoma (late submission)

Human: 4 Mb finished, 2.8 Mb unifinished
Mouse: 1.2 Mb finished, 305 kb unfinished

av. Contig = 220 kb, longest is 1.45 Mb

no clone/mapping capability yet, collaborating with others on 22, with Evans on 11 and possibly 15.

Critical issues: space (has commitment for new space from university); lab organization (everyone takes project from beginning to end), question is how scalable – currently can do 3Mb/yr this way, plan 7 next year, 12 the following year, have relatively modest scale up plan up to 40 by 2004; need more mapping capabilities globally, will continue to "depend on my friends," hope that there will be some groups that will have excess mapping capacity to supply others.
Other issue is quality assessment, how to come up with community-wide agreement of how to measure 1/10,000 and do that consistently. Need operational definition, not a specific standard yet. Have raised default settings for phred/phrap

Livermore/JGI

4.2 Mb fnished, 3.8 Mb submitted (difference is verification vs. restriction digests, analysis of the sequence)

average contig length 50.2 kb per clone (average); policy is that mapping contigs is 1 Mb, largest submitted contig is 1.02 Mb

error rate calculated to be 1.53 errors per 10,000 bases (data are binned by phrap bins, 97% are in bins between 40 and 90; sum phrap values over every base which is where the 1.53 comes from)

clone resources: have 168Mb of mapped clones from 19, 5, and 20. Have sufficient clone resources thru 9/99. Trying to keep maps a year ahead of sequencing.

throughput goals for this fiscal year is 20 Mb, for next FY is 40 Mb, and for FY2000 is 100 Mb (next Bermuda year hope for 30 Mb)

critical issues: staffing and training (80 hires in first year of operation, are behind schedule because people (particularly finishers) aren't there)
scaling up 10X this year
currently operating as virtual center
transitioning different strategies into a single process and infrastructure
keeping costs as low as possible in a start-up year

facility is 57,000 ft^2 in two equal size buildings, first should be ready in late summer; at least 80% of production sequencing will go to new facility. Production capacity that remains at labs will be to support technology development.

AR: don't believe 10X scaling

UTSWMC

4.35 Mb submitted

ca. 1/3 is earlier submission of random cosmids, rest are BACs and PACs which are directed to generating large contigs (2 Mb); largest contig currently is 804 kb.

putting out 800-1000 kb/mo now, at 1/10,000 by phred/phrap score. Anticipate completing a total of 12 Mb by August

finishing strategy now is based on availability of cheap primers, strategy moves most clones into almost finished condition (2-3 contigs), many of which have very hard regions

about 6.5-7Mb unfinished

clone resources: de Jong PAC or BAC library, mapping rate exceeds sequencing. Have several 2.5 Mb sequence-ready contigs; bottleneck is finishing, so mappers double as finishers.

Critical issues: funding
Project scale to 20 Mb next year, double that the following year, steady state at 40-50 Mb/year
Second issue is identifying and training people.

ACGT-ABI (Ellson Chen)

5.0 Mb finished and submitted (60% of that came in last year)
200 kb average contig size, longest is 700 kb
check data by Phred scoring test, checking overlap regions → error rate of <1/10,000

critical issue: building up infrastructure, ideal rate is 2X/yr
retention of people is a problem, turnover is about every 3 yr. Currently most experienced has been around 1.5 yr, most 6-7 mo

realistic expectation: ca. 10 Mb/yr at about 40¢/bp

mapping is a problem, since depend on others for supply

current funding is ending in June, center is dismantling. Looking to private contracting funds, not sure what to do next.

Jena

Unique genomic human sequence:	5.29 Mb
Total	7.3
Unfinished in progress	6.9

In last year, did 3.7

Av contig: 280 kb, longest ca. 1 Mb

Mapping/clones – supplied by others (Berlin, Toronto, Korenberg, Poustka), as well as at Jena; want to build mapping capacity in house

Quality: checked 2 Mb overlaps, better than 1/10,000, have some 500 kb contigs with rates of 1/70-80,000. Don't have funds for resequencing

BMBF funding for human sequencing is committed through 2000, (UH: will be a second phase of human sequencing, question is what regions?)
Comparative sequencing – currently sequencing a 3 Mb region of mouse X
Also sequencing 4 Mb/yr for next 2-3 yr of Dictyostelium

Critical issues: space

Next year target is 9 Mb of human, 15 the year after (not including mouse, Dictyostelium)

BCM

6.5 Mb finished, 5.9 submitted to GenBank (difference is in annotation); 4.6 in last year, have about 2.3 Mb in closure, 3 Mb in random, and 3 Mb in libraries

average submitted contig size is 130 kb, longest is 650 kb. Have active "joining" activity.

Quality is better than 1/10,000

Maps: have produced about 12 Mb at BCM, Wash U; future of mapping source is through collaboration with Kucherlapati on 12 (24 Mb next year), and Sue Naylor on 3 (17 Mb next year), plus retaining capacity at Baylor for producing 7 Mb of sequence-ready map.

Will come in at about 7-8 Mb this year. 7/1/98 to 7/1/99 projection is 30 Mb (which is a more conservative calculation than last year's projection). Have gone from 8,000 reactions/mo to 40,000 (which has been achieved in the last month). So will actually have done more than the predicted number of reactions. Assumes doubling of resources.

Critical issues: institutional support has been good (will occupy 11,000 ft^2 by July and have commitment for a second floor next year)

TIGR

(Didn't get back from break in time)

Whitehead

7.0 Mb finished, 6.6 submitted (difference is in "sign off"); virtually all of this in the last year.

average contig is typical BAC size, largest is ca. 400 kb

clones/maps: had steady supply of sequence-ready clones, but they came in relatively small contigs, which is a problem (particularly in terms of validation). Have switched strategy to isolating probes from chromosome 17 YACs to enter megabase regions, haven't gotten so far as megabase-sized maps yet.

Current rate of shotgun is 1.2-1.5 Mb/mo finishing at 0.8 Mb/mo. Want to double the finishing to match the current shotgun rate
= 13 Mb in next Bermuda year, 18 Mb in next grant year

critical issues: have been making many changes as we try to understand how the whole system fits together; had made locally sensible, but globally crazy decisions, now in the process of developing a much more reliable, sensible pipeline. Shotgun and prefinishing are fairly smooth, issue is to increase efficiency. Finishing have a lot to do. Come out of prefinishing with 2.7 gaps/100 kb, of which 1.7 are covered in clones, but 1.0 of which is uncovered.
Space should do fine for a while, institutional support has been good. Issues include proper way to account for cost issues in economically sensible ways.

All template preps and sequencing reactions are prepared by 4 people using robotic systems.

Washington Univ.

Finished is 24.5 Mb human, 22.6 submitted (difference awaiting analysis and annotation); ca. 20 Mb in the past year
Average contig is 175 kb, largest is 967 kb.

Another 35 Mb is through shotgun and awaiting finishing, another 9 Mb in shotgun, another 19 Mb in libraries awaiting shotgun.

In QA exercise, had 6 errors in 200 kb.

Maps/clone supply: finally at the point where mapping can exceed sequencing, right now at about 2 Mb/wk of human sequence-ready clones. Using STSs to pull clones and develop rigorous contigs around them. Does suffer from lack of long-range continuity. Addressing that by getting more probes from YACs, initial results in terms of closing gaps by this strategy are promising. Have several contigs of 2 Mb or greater.

Have converted completely to de Jong library in terms of mapping.

Still unclear whether single vector and cloning systems will be sufficient. Will YAC library made with appropriate informed consent be needed?

Will finish C. elegans in the coming year, will increasingly shift resources to human which in their hands is easier. 9/1/98-7/30/99 projecting 60 Mb, then 100 and 130

Critical issue is pipeline. Need to make every step more efficient. On-going issue is personnel/human resources. Intense competition for talented people at all levels. Have to bring new people into the field (need long-term vision beyond 2005 to make this a real science; money; creating an atmosphere where this is a valued activity; having visitors come through and make clear that what people are doing is recognized on the outside).

Sanger

Total human finished 36.02 (22 Mb in the last year), 34.85 submitted

Average submitted contig length = 1 clone
Largest submitted contig length = 600 kb

Quality – internal exercise → 4 errors in 1.1 Mb representing 4 different organisms, all were incorrect base calls.

Coming year target is 40 Mb finished, 80 Mb through total pipeline (would be 80% of Sanger total); like to go to 80 in the following year, and then 100 after that.

Map/clone supply – have 280 Mb of contigged clones available (200 Mb in past year, largest contig is 12.5 Mb on chromosome 22). Asking mapping group to feed sequencing groups at a rate of 4 Mb/wk (total shotgun capacity)

Have used BACs to close several map gaps that were extensively searched for in YACs

Working on development of automated finishing systems, to automate selection of finishing reads, autoedit, reisolation of clones to resequence.

First afternoon – session A: Quality/cost

Purpose is to try to come out of this session with processes to come to terms with the issues of quality and cost

Francis Collins:

4 A's: Accuracy
Assembly
Affordability
Accessibility

Costs – December meeting was sobering in terms of reality of scaling up and reducing cost at the same time. Now thinking about cost remaining the same for a couple of years, and then it will necessarily have to ramp down if we are to complete the human and get going on the mouse by 2005. Also have to allow for the possibility of failure of a sequencing center.

FY97 – allowing $70M; next year President's budget allows for ca. $80M. NIH budget looks optimistic for the next few years, at least in terms of talk.

Wish we knew more about the relationship between cost and accuracy. Is it a steep curve, or a relatively shallow one? There are still people out there who are not convinced that the 10^{-4} accuracy goal is not necessary.

EL: what kind of data would you need to answer the question.

RW: talking about the base pair quality is not the issue. The greater challenge is in getting the DNA "that is not representative"{?}

JS: not worth measuring in that simple form (cost/bp vs. accuracy). Having sequence of the quality that assembles well gives the kind of accuracy we're looking for. gaps is a separate question

Can argue about the effect of allowing the introduction of controlled uncertainty in certain areas.

Description of 2nd sequence quality assessment exercise.
NHGRI will do again next autumn, "mixed" enthusiasm for doing another one between now and then

EL: Whitehead results
10 errors in 200 kb

- 3 "bookkeeping" errors – penultimate file submitted

ACTION: standardize, establish preventative system

- -- using GAP4, which uses majority vote (could be set at 100% editing, but wasn't); actually the right data were then, so had made them think about ways to approach
- 2 "GC" vs "CG" – Whitehead interpretation still don't agree with checkers, don't understand why
- 2 deletions in local region
  - 252 bp deletion occurred in cosmid
  - 600 bp deletion in single M13 used to span
  - deletion in large restriction fragment

  ACTION: software, more & more rigorous change in processes

Goal of QA is to improve process, which is how the QA exercise was used at Whitehead.

BR: 4/5 clones chosen had already been reassembled by their new procedure at higher phred/phrap criteria. They came back with 3-7 discrepancies, which had to do with stringency differences used. Have resequenced the discrepancies directly off the target clone (BAC or PAC).

Fifth clone was one of the "clones from hell" which was old clone that had not made it through the new QC process yet. Has convinced them to do all of the confirmation in house.

EL: important point is that different groups have different finishing/submission rules. Would it be useful for procedures to be shared, to encourage assessment of "best practices." What about small working group to deal with finishing issues, e.g. share data, experience, etc.

RM: would like to know about group experience of rearrangements in subclones. We've found 2.

RW: don't know frequency, know that there are several sources of problem (ligation, growth, tracking). Not a high frequency but they occur → policy of not relying on single clone, which solves the problem for M13 subclone but not for cosmids/BACs

PG: have seen cases of recurring specific deletions in M13 subclones

RW: could have arisen in cosmid growth, giving mixed population. Have seen cases in which you look at the restriction digest, you can see a faint band.

JS: experience with tandem/inverted repeat in M13, which eventually was dealt with by subcloning in pUCs.

RW: have examples of deletions in pUCs, although it may be less frequent.

BR: all our sequencing is done in pUCs, almost all of our discrepancies can be traced back to sub-population of deletions in cosmid. Approach to dealing with this is sequencing directly off large clone.

RW: is the frequency of mutations in subclones low enough so that it can be ignored if you set the policy that each base has to come from two different subclones?

PG: feeling is that it is rare to have a mutation or deletion confirmed by a second sequence. It's a reasonable policy.

EL: are these getting buried in lab notebooks, or should they be collected for analysis?

DC: if it's random and doesn't have biological information, then I don't care. But I won't know if it's telling me anything new until we collect them.

PG: have looked at our own data sets pretty thoroughly and there are lots of anomalies that show up (incl. various kinds of contamination)

MV: Sanger QA exercise. Didn't reassemble. Exchanged data among teams and relooked at data. Four errors were detected, all were incorrect base calls. One was a misinterpretation of a weak G after an A (a problem inherent in the use of terminator chemistry); a second was due to an autoedit error in a lower quality region (Action 1: autoedit now prevented from editing in a region of two reads only, action 2: randomly recheck older clones for possible error)

Conclusion that checking within the different Sanger Centre teams will continue to ensure consistency.

EL: ask about shatter libraries, which you say help to deal with difficult regions. Can we get systematic analysis of what/how was helped?

RW: doing that.

EL: what about continued/expanded QA exercise?

FC: repeat what we've done, but open up to all comers?

EL: hear from groups who aren't obliged to participate

JS: would like to. Two points, want to be part of organization that plans it and sets the rules

EL: what would you do differently?

JS: not in favor of distributing materials and doing resequencing. Maximizes value if the resequencing is done by the lab whose errors were detected.

EL: didn't find it to be that much work to do. Fingerprinting was crucial.

RG: speak in favor of reagent swapping. Sounds burdensome, but was remarkably straightforward. There were issues of uncertainty in the electronic data that were more of a problem.

EB: JGI enthusiastic about participating. Comfortable with the format

YS: would be happy to have sequences checked, not clear that they want to do checking.

GG: French are at beginning, would be interesting for us as we have opted for another technique (LiCor machines).

EL: unanimity

PG: one thing that wasn't checked was clone fidelity. Would vote for exchanging sets of clones overlapping the one whose sequencing is being checked.

EL: at the moment don't even deposit the evidence that the clone is faithful.

RG: what to submit in terms of quality measures.
Propose - estimate of error rate for submission
- distribution of BCM-Phrap <40
- list of consensus changing edits
- list of lowest quality bases

JO: GenBank has done the coding to incorporate the phrap values into the submission, as well as some other things. Some people have it already. So it's possible to put that information into the record.

GC: will work for the data that GenBank collects, other databases collect more

DC: go back to fidelity issue. Any method for validating has a particular resolution to it. We've been talking about methods for validation but not sensitivity to different types of errors.

PG: can't say anything about MCD mapping of BACs other than it's clear that the data are harder to interpret because of band overlaps and fewer ends.

RW: our data say that fingerprints give measurements that are good to one percent

EL: another working group on genomic fidelity.

Now let's move on to cost

What are we trying to accomplish? There are multiple reasons to measure cost, e.g. find out whose cheapest, to do process improvement, cross-comparisons to do best practices comparison, to do projections. Process improvement and projection are the two most important.

In previous discussions, obvious that definitions/terms/etc. are different. Want to be able to have pair of producers be able to sit down and have a sensible discussion.

CM: first don't try to reduce it to a specific number.

EL: another working group, include NIH, Sanger Centre cost accountants.

JR: sounds horrendous in terms of amount of time involved, would be willing to talk about principles.

BR: key issue is to be able to learn from each other in a safe context that is not being forced upon us by the funding agencies. Want to share common experiences so that we learn how to share experiences.

FC: trying to move into a more cooperative mode. But if we don't move along the cost reduction curves, we will be in trouble. Can do this individually or can try to learn from common experience. Not to do this seems fairly difficult to defend.

RM: can't get costs down without understanding them.

EL: cost drivers. Where is the money now?

RW: half personnel, half non-personnel
In non-personnel, first is sequencing reagents, rest is lots of little items (glass plates, service contracts, plasticware, finishing reagents, foil tape).

JS: 30% salaries (30% finishing, 30% production, 30% ancillary activities), 50% reagents, rest is overheads

EL: sounds like major issues are decreasing labor and decreasing reagent costs. Won't be able to drive down unit reagent costs, have to focus on reducing volumes

Rick wilson will chair another working group on consortium buying.

RW: to reduce volume, can see how to reduce volume by half. Would seem to need new technology to go lower.

BR: not cost of reagents, but efficiency

PG: haven't discussed setting standards for continuity, think we should. 10^{-4} standard helped to drive people to meet it, should do that for continuity also.

EL: in December decided it would be useful to report it, and set goals rather than setting standard. What's a realistic goal for a year from now for average contig size?

PG: 500 kb is reasonable.

EL: weight average size is probably relevant

PG: average gap frequency in a region of DNA should be one in 500 kb.

RW: gaps are what takes the work to fill.

EB: also issue about how gaps are characterized. Number alone is much less information than being able to describe size and orientation of contigs around them.

RW: these are interesting points, would like to agree on first number before going to additional descriptors.

Issue is one of deferred costs. Don't want to give misimpression of how much it is going to cost to sequence the genome.

PG: same issue on individual clone level.

JS: isn't clear that it is more economical to close each gap as we go along. Don't think we should be constrained by this kind of thinking

First afternoon – Session B: Data Release & Availability

Jim Ostell - three versions for reporting sequence + the more complicated fourth version (subsets A, B)

EB: what are the reasons against the most complicated version, should be relatively easy to do and publicly available. Embarrassment that we can't do it now

DC: historically these have failed in the past, not because they weren't a good idea, but because of the way they have been implemented. Need standard format

JO: basically offered versions to see what is tolerable to people

CM: can you support different versions for different groups?

JO: yes

BR: much of this is already available

JO: going to different web sites and clicking on things is different from having a single site in a defined format

BR: what's wrong with the file you've already got

RD: good idea

JO: our purpose is to ensure that what is in GenBank is what the centers think is in

EL: implementation worth discussing, let's just try to get it as right as possible. JO should write it up, circulate it, get comments, test in trial period before going full blown.

JO: agree and disagree. Would recommend version 3 for the sequence contigs and then continue to discuss the map version.

GC: agree in principle, databases have to agree on what will be exchanged.

AR – Web Survey
(waste of time)

RG: data release policy. Is there compliance?

AR: some centers do not release unfinished data to their Web pages, only finished data.

TC: only one of JGI sites is currently; other two sites are coming up very rapidly

FC: preference is to have the unfinished data in the public database, rather than just on the Web site.

GG: in France, at very beginning. Have both in-house and collaborative projects. In case of in-house projects, will adhere to Bermuda standards; as for collaborators, dealing with on a case-by-case basis. Until recently, majority of collaborative projects were not human. In the near future, will be human EST project. Human genomic sequencing will be on chromosome 14 and 3; will adhere to Bermuda standards.

DC: understand that previous statements from this meeting is that large-scale sequencing centers operative under the rules that all sequence is released, regardless of collaborators' wishes.

RG: take message back to French sequencing center and that this group can help exert pressure on the ministry/government

Japan?

Kikuchi: JST only requires finished data, not primary. Release of unfinished data is at discretion of investigators. JST has no intention of delaying data release [but only does so every three months]

Sakaki: principle of the JST is to release finished data every three months to DDBJ. Finished data is sent to JST every three months, and at same time to Eleanor Roosevelt Web site. Our group is releasing unfinished data nightly. Will maintain policy when move to Riken.

FC: average time that JST holds data? What is JST doing with the data, why does it want to have the data?

Unresolved discussion about the fact that JST policy does not adhere to Bermuda principles.

Second Morning, Session A: Sequence Claims and Etiquette

AR: First rule "mapping investment does not automatically entitle sequenced claims over the same region until a sequence-ready map is generated" needs more definition.

JM: chromosome 2 recently started. Taking a chromosome-wide approach to the mapping. Mapping rate will exceed the 3X allowed by last year's rule.

DC: plan? You are using already mapped STSs to pull out BACs to start, how will the sequencing choices be made?

JM: can sequence on a regional basis

RW: fact that we are or will sequencing from a number of sites creates a practical problem in terms of reporting to the HGSI. Clearly mapping exceeds sequencing capacity.

DC: I understood the one year thing to be related to the sequencing, and not to pulling clones across the chromosome.

EL: no one wants to limit sequencing, concern is to avoid people restricting access of others to regions by holding claims. Contiguity has a lot to do with his, if you are holding regions you should bring those to contiguity (closing) before moving on to another region

We're not on the list. We would have put on a claim for 17 ptel to qtel. We've now decided to focus on a 20 Mb region first, and would like to move on to other regions of that chromosome after finishing it. Right now we have 70% of the chromosome in BAC contigs but they're scattered.

AR: don't like whole chromosome claiming and restricting territories. Many groups around the table have no mapping capabilities and are dependent on larger centers for clones. Puts smaller center at a disadvantage. Should go for smaller claims and put more energy into finishing contiguous regions.

Our claims are on our Web page. They are very small, typically 1 to 5 to 10 Mb. not claiming all of chromosome 21 because there are many groups that have come to agreements. In other places, have found it very hard to come to agreement with American centers.

I will put up regions that have already been claimed, e.g. on chromosome 7

DC: this list doesn't preclude additional claims, but it should involve discussion among groups involved.

EL: it's not the details of the rule, but it's the clear commitment to finish the job

RD: can separate continuity from claims for whole chromosome. Has to do with planning and funding commitments. Goal is to get complete sequence from end to end, and will have to be done in a systematic and complete way. but right now, we can only work in much smaller units. This is a five to seven year job, so a one-year time span for commitment seems to be a little bit of a disconnect.

DC: cafeteria metaphor. We've not been to this cafeteria before. Will require experience and TRUST. Legislating it won't work.

DB: we have made commitments to getting whole chromosomes done. Doesn't mean that others are precluded from working on it. We retain the commitment to ensure that the regions that others don't do will get done.

Not all Sanger claims are in there because of mechanistic difficulties of getting the information in. e.g. on X, the markers agreed on by the X workshop are not in the Index yet.

DC: there are some situations in which the boundaries are defined to the base pair and in which lots of negotiation have been done.

{long discussion, including individual centers plans, didn't take notes}

DB: new proposal

Difficulties in using site, particularly because of marker selection

	AAFM	AFM	
	AFM		AFM
AFM	AFM		AFM
			H2A.1
	AFM		AFM
	AFM		AFM
			D20S492
AFM	AFM		AFM
	AFM		AFM
800	2000		3000

Also, ideogram showing status of claims (Human Genome Map Status); on Sanger Web page – called map status, shouldn't be confused with accomplishment, only represents claimed activity.

URL for HGSI (http://www.ncbi.nlm.nih.gov/HUGO/)

Additional markers: David Bentley and David Cox have volunteered to act as checkpoint and to put those markers onto high resolution RH map.

Set up e-mail list for people to submit their specific sequencing plans for the next year. Also try to expand the HGSI to distinguish between mapping, construction of sequence-ready maps, and regions of active sequencing.

Second morning – Session B: Sequence-Ready Maps & Resources

Phil Green: table showing differences between last year, this year, and claimed accomplishment for this year.

		Feb 97	Feb 98	Actual 97		Claimed		Predicted		
								97		97
Sanger	14.6		36.0		21.4		22		35	
St. Louis		4.8		24.5		19.7		20		24
Whitehead		2.1		7.0		4.9		7.0		20
TIGR		2.7		6.4		3.7		4		11
BCM		3		6.5		3.5		4.6		12
Stanford		0.3		0.65	0.35			5		
UW		0.59	3.65	3.06	3.2		6			

Chen-ABI		2.4		5		2.6				3.5
DOE		4		4.2		0.2				20
Roe-Ok		3.8		4.1		0.3				5.5
Jena		1.5		5.29	3.79			6		
UTSW	1.6		4.35	2.75			5			
Sakaki	2.7		5.1		2.4				3.4	
TOTALS		44.1		112.7		68.6				158

"Optimism Factor" = 158/68.6 = ca. 2.4

A. Resources

a. BACs

Availability/upcoming libraries
Criteria for quality

b. YACs

Mark Adams – CalTech status
Library D – (see overhead)

(discussion of TIGR/Washington experience to date with BAC end sequencing) end sequences submitted to GenBank weekly, Web site updated nightly. Both TIGR and Jena have already used BAC ends to select clones for sequencing

Peter de Jong
List of libraries prepared

Quality assessment for libraries
Look at 1% of clones for insert size (currently doing ½% which may be enough)
Probe with at least 10 unique probes (probably will do more)
Detailed contig map using STS markers spaced at regular (5-10 kb) intervals; currently have done one for chomrosome 14 contig
Possible probing for α-satellites, transposons

EG: is filter supply capability adequate?
PdJ: can fill requests w/I one to two weeks, also provide UK, German resource centers, as well as Research Genetics

EG: is there are need for PCR screenable pools?

EL: too many libraries
RM: quality of commercially available pools was poor
EG: what about one or two?

EL: are there labs out there who want to screen for regions who would prefer pools to filters? That may be choice of positionally cloning labs, so there may be need for one or two

DC: PCR gives noise. Even for small groups, filters followed by local PCR screening is good. No need for pools.

EG: what about quality assessment criteria?

AR: is there a need for BAC libraries with even larger inserts, 400-500 kb?

BR: YACs, bacteria being assembled so that isn't the issue.

PdJ: could do 300 kb, but electroporation is limiting. May be other approaches to getting larger DNA into the bacteria. Confident that once the DNA is in, it can be maintained in bacteria.

AR: like to go on record as wanting to have at least one library with large, 300 kb BACs

DB: also in favor of large BACs; could be very important for closure

BR: what other libraries are available for comparative genomics

PdJ: there is a lot of interest in many other libraries (organutan, gorilla, Arabidopsis, fugu, etc.), issue is one of funding.

EG: other items, first YACs

Suitability of current libraries: probe generation, complete sequencing
New Libraries??
PCR Screenable??

PdJ: intend to use new vector with yeast centromere, His-3; could transform into yeast spheroplasts at same time as bacteria.

EL: will select for presence of human ars function, may bias library

PdJ: they are pretty uniformly distributed

EL: better to put yeast ars into vector.

DB: is there enough DNA from approved sources to make new YAC libraries?

PdJ: yes

EG: characteristics of new libraries?

EL: for probe generation, want largest YACs possible, so that they will cover large regions. For sequencing, probably want them smaller, something like BAC-size YACs to have the human DNA in another system.

EG: Mapping strategies

Optimal starting STS resolution
Experience and cost for typical chromosome
Metric for monitoring progress/quality

DC: people can define by what they mean by "a lot' in U.S. goal was average of an ordered marker every 100 kb. There are 30,000 but they are on different maps so that we can't tell how good the order is. What are you trying to do with the 30,000? Continuity or coverage? current maps aren't at 100 kb resolution. How do you use that to pull clones in contigs? Taking STSs that are known to be separate on RH maps, which will give coverage, not contigs.

On chromosome 4 with 1000 ordered markers, pulling BAC for each would give over 85% of chromosome covered. Need higher resolution map (50 kb) to get contigs.

Take home lesson is that taking all markers is less efficient in getting coverage.

Available maps allow thoughtful picking of BACs to give coverage, not continuity

DB: different approach. Going for density of 15 markers/Mb, don't know yet if that will be adequate. On GB4, that gives about 1 Mb resolution, don't order them. Use to pull sets of clones and depend on fingerprinting, STS content to do ordering. Results so far are encouraging, generally get about 60-70% coverage.

EL: coverage isn't the issue. What about continuity?

DB: the approach is directed to the initial phase.

EL: not clear what resolution means. Issue is how to get enough markers to almost cover the chromosome, how to get the rest of the markers to close.

Real problem is how to get markers where we don't have them, how to identify closed regions where we have them but don't know

Old theory says that a marker every 35-40 kb will give one to zero gaps per megabase. That's what we will all eventually have to do.

DC: resolution will tell you up front where the gaps are before pulling all clones. Question is what's the most efficient and cheap way to do it? We think that high resolution RH map is the best way.

EL: density of markers will have to be 3 per 100 kb no matter what the initial strategy is

DC: good news is that initial indications is that so far the marker generation looks pretty random.

EL: ESTs show biased distribution

DC: that's the reason we aren't depending on ESTs for markers

JM: worry about gaps later

EG: move on to idea of "typical" chromosome; are we at the point to be able to evaluate costs

EL: cheap to do first 50%, don't know what it will take to do first 99%

DB: can't estimate cost of finishing chromosome until last gap has been closed; impression is that cost of closing maps is in excess of production; on chromosome 22, have made extensive use of YAC map

AR: re BAC sequencing. Had thought of that project not only for gap closure but as mapping tool. What is attitude of mapping groups?

DC: the way to get high confidence maps is to use as many pieces of independent information as possible; likelihood of using only BAC end information to get high confidence maps is very low, but that information will be useful in combination with other methods. so now, when we make new random STSs, we make them off of BAC ends.

MA: in Arabidopsis, there is an end sequence database and it is the sole source of clone information. In human the situation is more complex, and agree with DC that multiple approaches are necessary.

EG: move on to metric for monitoring progress/quality. Does everyone mean the same thing by sequence ready? Is there common experience?

EB: current consensus of 10-15¢/bp?
(not concurrence)

JM: last year, mapping was 7% of total cost (will go up as mapping gets harder)

EG: map validation

DC: working group on clone validation, validation to genome will take this on

EG: centromeres? How close are people going ?

DB: have one on-going collaborative study where there has been an extensive study on chromosome 10 that has detailed long-range mapping information from pulse-field gels, etc. Large component of work is validation of clones and maps.

AR: telomeres are issue also.

BR: on 22, it wasn't as bad as Moyzis' stuff on 5, could tell when we were up to the telomere.

Second afternoon – Session A: Annotation

Waste of time discussion.

Second afternoon – Session C: New Technology

Rick W: Modification of the ABI 377 (Tibbets)

Concept: Increase the pixel density per scan to enable an increased number of lanes per gel

Current status:

- modifications made to two instruments at the GSC
- Exportable ("plug" & "play") version of hardware completed'
- Currently working with 96 lane gels to improve reliability (problems have been with pouring and loading gels reliably)

Talking with Baylor, U. Washington, TIGR about exporting

Have not addressed the issue of the economic impact of this 'improvement'

RG: the real issue is whether there is any lowering of data quality at that lane density

LH: half of production labor is taken up with pouring, loading, breaking down gels

RW: there are other advantages even if increased lane density doesn't help (e.g., possibility of fifth color)

EC: ABI 96 lane development; have been testing for less than two weeks, looks pretty promising, but too few gels to evaluate consistency. Should be available commercially some time in the next quarter

Capillary arrays:

Chris Martin: Joe Jacklavic development

96 capillary module, sheath flow detection

4% linear polyacrylamide, 2 hour run

in "developmental testing" phase, calibration and software development. Have had 23 runs of production sequencing samples

evaluating data quality

so far, in first 400 "pretty nice and consistent", steeper fall off>450 than ABI

have to evaluate the up front effort/cost of loading/maintaining the capillaries

uses confocal detection which is protected by Mathies patent licensed to MD

RW: experience with MD Megabace machine

First few months plagued with problems, have to relearn how to troubleshoot; capillaries need maintenance, matrix can be problem

In last few weeks, have gotten more reliable runs out to 400-450 bp

Still issues of capillary coating, lifetime. (so far has been variable)

MA: only looked at plasmids and M13, aren't quite as far as long as GSC

[lost some of the discussion of capillalry instruments due to computer problem}

Tony C: microchannel device,going

Continuation of Bermuda III notes.

Up-front automation –

Chris Martin: get integrated system based around 96 tip Hydra pipetting head
Using in production to set up cultures, frozen stocks
Developing for template preparation, sequencing reactions
Starting to have an impact in production

EB: collaboration with Lander to co-develop front-end robots (to be built by IAS)
All in planning now.
Main effort is in template preparation

GE: two large Sagian robots using large robotic arm to feed Hydra pipetters and MJ PCR machines. Arm is reliable, have had problems with PCR machines.

RG: have imported one, but have chosen to do the thermocycling off-line. Have expanded other capabilities. Philosophically is important to stick to one platform once it's been started.

JR: back end system at Sanger for automating pre-finishing step; robotic arm system that transfers 96 well plates → 384 well plates for storage, and another tip for rearraying

Chemistry/biology

New enzymes coming along. Amersham is working on mutants to FY enzyme; might end up not with a single enzyme that deals with everything, but a toolbox of different enzymes that are good for particular problem regions.

AR: making mutant polymerases that can read through problem areas.

EC: have tested 4-5 different enzymes in the last 6 mos. haven't had any consistent results at this time.

Dye chemistry

RG: are using BODIPY dyes exclusively for dye primers. Terminators don't look promising at this time. Patent issues remain complicated. It's only threat of repercussions from ABI that is preventing commercial distribution at this time.

Also mentioned Bob Weiss low copy number vector that can handle large inserts (so far 12-15 kb) with runaway replication capability.

"shatter libraries" to fill gaps. Take PCR product or restriction fragment, sonicate, purify small fragments (100-500 bp), clone into M13, sequence and assemble as a small shotgun project.

Has worked on every gap tested so far.

BR: we published this a couple of years ago and have stopped because it's easier and cheaper to use BIG dyes to sequence off target clones. It's always worked for us [Hasn't always worked for others].

AR: what about chemical sequencing?

RW: just hasn't been all that clean, this works better.

Eventually, want to develop Finishers' Toolkit.

BR: love the Wash U double acetate procedure to give good template quickly (on the Wash U Web site)

EB: George Church project with mass labels, gives 400 hundred labels instead of 4; in a reasonable stage of development

AMS '98 in St. Louis, 10/8-10

FC: conversation leaves me uneasy. Level to which new technology is getting into the system is less than overwhelming. New technology is being worked on, but doesn't seem to be connecting to the people in this room. How can technology insertion be improved?

Ed Southern: don't need to be so pessimistic, there are lots of development projects out there.

FC: understand, but still wonder whether we have the proper connections set up.

RM: worst time to try to do this is when you're in the middle of a big transition to higher scale up.

Second afternoon – Session C: Next meeting

FC: what happens now section. Start with working groups proposed yesterday. Have identified four proposed chairs:

Finishing – Waterston, chair
P. Green
Someone from Sanger
Lander

Chen

Mandate a little vague, and group will have to define. Should include comparison of different approach, database of difficult gaps, some other things

Clone fidelity, genomic fidelity – Cox, chair
P. Green
McPherson
Birren

Cost accounting – E. Green, chair
J. Rogers
Washut (?)
Others

Goal is not to figure out who's cheapest but to figure out a way how to dissect out the costs.

Consortium buying – Rick Wilson, chair
Myers'

Others interested should identify themselves to the chairs
Out of these groups will come much of the agenda for the next meeting

Another issue not addressed yesterday is mouse sequencing and data release. NHGRI centers are being encouraged to devote as much as 10% of their effort to mouse. What about data release? Why not adopt same policy as for human?

DB: agree, policy derives from worm, not species specific; objective is to build community.

ES: why restrict to mouse? Apply to all organisms being done on a genome basis.

AR: DFG applying that policy to the two models it is funding.

FC: metazoans. Say that Bermuda principles should be applied to other metazoans and urge other sequencing projects to adopt these practices.

RW: that plus leading by example

FC: bring to closure – not much unanimity about making a statement about procaryotes.

RW: do feel strongly, will release. Ditto for Sanger, Roe

TC: another part of DOE is sequencing pathogens and policy hasn't been decided. And maybe will decide not to release "for reasons of national security."[!]

PG: seems strange to exempt any organism. Should be all-inclusive

MB: what this group has is serious scientific experience of the value of doing this. That's an important message the group has to offer.

YS: I'm in a difficult position to be able to agree (although I do personally) because in Japan there are many agencies with different policies.

ES: the statement should reflect that these are the personal views of the people at this meeting.

TC: does this apply to full-length cDNA sequencing?

BR: same urging

FC: don't see why not

Think what the sentiments of the group are is that we believe that the value of genome sequence/full-length cDNA/large scale information is sufficiently great and immediate that we think that it should be released immediately into the public domain, that we will adhere to these principles, and urge others to do it.

Next year. Need to have a gathering of large-scale sequencing enterprises once a year. Here, or another model, e.g. alternate between UK and US near gateway.

DC: Bermuda becomes less and less likely to show up. Venue should make it easier for people to show up.

Vote 19 to 7 for alternating.

YS: HGM '99 is in February in Australia.

FC: content. Time to focus on particular topics. Hope that working groups will be good jump start to sophisticated discussion about those topics. Other thoughts about agenda.

AR: who will be invited? Should define what a large scale sequencing center

From:	Collins, Francis
Sent:	Sunday, February 15, 1998 11:32 PM
To:	Guyer, Mark; Peterson, Jane; Felsenfeld, Adam
Subject:	FW: bermuda III

FYI. I don't find John's response very reassuring. What do you suppose the quality of the Sanger Centre sequence is?
FC

-----Original Message-----
From: **Collins, Francis**
Sent: Sunday, February 15, 1998 11:28 PM
To: 'John [redacted]
Subject: RE: bermuda III

Hi John,
Your response surprises me! I would have thought that this is exactly the time to discuss the quality assessment issue, since there is actually some data in hand (the second quality assessment exercise we have been carrying out). I am hoping that at Bermuda we can get a real sense of what the curve of sequencing accuracy vs. cost/bp looks like, and also an idea of how difficult (and expensive) it is to close gaps. As of December, we are moving toward a plan that finished sequence from NHGRI centers should stretch across 500 kb or more, but there is not a lot of data about how difficult that will be to achieve.

Similarly with costs - I was hoping that we could go well beyond the dollars in/base pairs out analysis, and try to get a more detailed sense about the origins of the costs per lane and lanes/finished kb. That would seem to be the best way to figure out how we're going to get costs to go down (if they are going to). I agree that diversity of approaches is a critical feature - but how we will all benefit from the diversity of methods for cost savings if we don't do this analysis? All of that will take some time.

I certainly agree that sharing and data release are critical, but I am
under the impression that the past two Bermuda meetings rather set the standards
here, and that most centers are now adhering to them
I'd be interested in the thoughts of the rest of the observers on this e-mail network.
Regards, Francis

-----Original Message-----
From: **John Sulston** [redacted]
Sent: Sunday, February 15, 1998 7:26 PM
To: [redacted]
Subject:Re: bermuda III

Dear Michael,

Personally, I'm not enthusiastic about spending extra time on quality and costs.

It's good that methods of quality comparison are being explored, but it would be wrong in my view to codify things rigidly at this stage. There's still a lot of technical progress to be made, and I for one don't want to see the product nailed to the ground so early in the game. The aim here at the SC is to produce the best possible product given the current state of the art, and at the same time to help push forward the state of the art. I don't believe we need a huge amount of time to deal with the matter.

On costs, we are all working at driving them down in our own ways. Diversity is one of the most powerful weapons we have, and our diverse technologies are best explored through the normal scientific meetings. So a brief statement from everyone is all that's neeeded, I think.

On the other hand, what Bermuda is really all about is the content of the next two sessions: sharing and data release. The issue of annotation is also becoming more important. So I don't believe that Saturday should be squeezed in the way proposed.

This is my personal, undiscussed opinion. I'll be happy to enter into a dialogue if others disagree with me.

All the best

John

Guyer, Mark

From: [redacted]
Sent: Tuesday, February 24, 1998 3:41 PM
To: Guyer, Mark; Collins, Francis
Cc: ostell@object.nlm.nih.gov
Subject: contig info

Here is the 3 versions of sending information on finished sequence that we discussed..

1) just a list of finished sequence
2) ordered lists of finished sequence by contig
3) same as (2) but also giving base pair coordinates and orientation to assemble without any additional computation.

I think they are self evident, but I can discuss more if you like.

I also include two proposals (below ++++++++++) for describing the whole map including unsequenced pieces. Obviously this is more complicated, so I had planned to bring along the overhead in case it came up, but to focus on the 3 simple models first (and maybe only).

Jim Ostell

>>>Version 1:

U00001
U00058
U11123
U55555
U44444

>>>Version 2:

Contig_name: 22ctg5
U00001
U00058

Contig_name: 22ctg8
U11123
U55555
U44444

>>>Version 3:

Contig_name: 22ctg5
U00001 1 100505 +
U00058 53 150333 +

Contig_name: 22ctg8
U11123 1 175033 +
U55555 44 165000 -
U44444 32 150551 +

```
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
Version 1:
==============================================================================================================
Contig Name: Chr_22ctg17
Accession: No
Title: Contig that Covers 22q13.3 on Human Chromosome 22
Length: 3500000
```

Subsumed Sequences: U51560 in Z80896, Z78421 in Z80773

Clone	Length	Accession	From	To	Orientation	Library
dJ117O20	100000	-	-	-	unknown	PAC
dJ116M15	23592	Z97351	0	23591	unknown	PAC
gap	unknown	-	-	-	-	
fF64F4	44516	Z82187	0	44516	plus	FOS
cN79E2	39683	U51561	14780	39682	plus	COS
bK358H9	100000	-	-	-	-	BAC
dJ67C13	56913	Z80896	0	56912	plus	PAC
cN73F4	38468	U51560	-	-	minus	COS
fF111A3	42776	Z79999	104	42775	plus	FOS

```
=============================================================================
Version 2:
=============================================================================
Contig:      <name>
Accession:   <accession>        [OPTIONAL]
Length:      <length-in-bases>
Description: <description>
Chromosome:  <chromosome>
CytoBand:    <cytogenetic band>  [OPTIONAL]
Clones:
        <clone> <from> <to> <lib-code>
        <clone> <from> <to> <lib-code>
        <clone> <from> <to> <lib-code>
Sequences:
        <accession> <from> <to> <clone(s)> <lab>
        <accession> <from> <to> <clone(s)> <lab>
//

-----------------------------------------
Library:     <lib-code>
Description: <description>
Type:        (pac/bac/p1/cos/fos/etc..)
Vector:      <vector>
Creator:     <creator>
Supplier:    <supplier>         [OPTIONAL]
Prefix:      <prefix>           [OPTIONAL]
//
-----------------------------------------
```

CHANGES TO PROGRAMME

- Dr Jean Weissenbach will not be attending the meeting due to a skiing accident. Dr Gabor Gyatay will attend in his place.
- The venue for the pre dinner drinks on Saturday evening will be the Penthouse Suite on the 7th Floor.
- The venue for the dinner on Saturday evening will be the far end of the Tiara Room Terrace on the Mezzanine Floor

ADMINISTRATION

- Please ensure to let me know if your return flight details have changed so I can keep the Ground Transportation Company fully informed.
- Please don't hesitate to contact me if I can be of any further help during the meeting.

Jilly Steward
Meetings & Travel Manager
The Wellcome Trust

FAX TRANSMITTAL SHEET

NATIONAL HUMAN GENOME RESEARCH INSTITUTE

National Institutes of Health
Building 38A, Room 604
Bethesda, MD 20892
(301) 496-7531

TO: Francis Collins and Elke Jordan

FAX NUMBER: 20837

FROM: Mark S. Guyer, Ph.D.

DATE: August 19, 1997

# of pages including cover sheet: 10

Return FAX [redacted]

[redacted]

Comments: This is the request from Wellcome that I mentioned in my e-mail yesterday. Jilly Steward is soliciting input for the attendance list at Bermuda III. It seems to me that most of last year's attendees should be invited back, with the exceptions of Ung-Jin Kim, Dick McCombie, and possibly Gert-Jan van Ommen. More importantly, I suggest adding Gerry Rubin and Lee Hood. I also think we should suggest one or two members of Jane's Advisory Committee (specifically Ira Herskowitz and/or Richard Lifton) and one or two additional members of Council (Bob Horvitz, David Valle, and/or Alan Williamson). I think that both of those committees would benefit from some more direct exposure to the Bermuda discussions. Should we quickly discuss this tomorrow afternoon?

THE WELLCOME TRUST

27-1

richard_lifton@ ~~[illegible]~~ qm. yale. edu

210 Euston Road
London NW1 2BE
[redacted]
E-mail: j.steward@wellcome.ac.uk

FACSIMILE TRANSMISSION

Lifton ✓
Herskowitz ✓
IRA@cgl.ucsf.edu

Rubin
Adv Council member (Herskowitz
~~Ielglem~~)

Council member - Valle
Hood
Horwitz
Williamson

Ref: JS/SO'D/Fax449

TO: Dr Mark Guyer

FAX NO: [redacted]

FROM: Jilly Steward

DATE: 12th August 1997

No. of pages including this: 9

Dear Dr Guyer

Further to Dr Michael Morgan's letter of 24th July, I am writing to you to ascertain whom you think should be invited to attend the 1998 meeting.

I attach a copy of a list of those who attended this years meeting and also a list of those who were invited but were unable to attend.

It would be of immense help if you could let me have your comments back by Monday 8th September when I return from leave so that she can proceed with the planning of Bermuda Three.

With kind regards.

Yours sincerely

Jilly Steward
Meetings and Travel Manager

Second International Strategy Meeting on Human Genome Sequencing

List of invitees who declined

Professor Michael Ashburner
University of Cambridge
Department of Genetics
Downing Street
Cambridge
CB2 3EH

[redacted]

Professor Martin Bobrow
Department of Medical Genetics
Box 134
Addenbrooke's Hospital
Hills Road
Cambridge
CB2 2QQ

☎
[redacted]

Dr Michéle Durand
Ministérc de l'Éducation
Nationale de l'Éseignement Supérieur et
de la Recherche
1 rue Descartes
75231 Paris Cedex 05
France

☎
[redacted]

Dr Richard Durbin
The Sanger Centre
Hinxton Hall
Hinxton
Cambridge
CB10 1RQ
United Kingdom

☎
Fax
[redacted]

Mr Kanji Fujiki
Director of the Life Science Division
Science and Technology Agency
2-2-1 Kasumigaseki
Chiyoda-ku
Tokyo 100
Japan

☎
[redacted]

Dr Eric Lander
Whitehead Institute/MIT centre for Genomic
Research
One Kendall Square
Building 300
Cambridge MA 01239-1516
USA

[redacted]

Dr Hans Lehrach
Max-Planck-Institut fur Molekulare Genetik
D-14195 Berlin (Dahlem) Ibnestrasse 73
Germany

[redacted]

Dr Frank Laplace
Federal Ministry for Research & Technology
Heinemannstrasse 2
D-53175 Bonn
Germany

☎
Fax:
[redacted]

Professor Kenichi Matsubara
Osaka University
Institute for Molecular & Cellular Biology
1-3 Yamada-oke
Suita Osaka 565
Japan

☎
[redacted]

Dr Richard Myers
Stanford Human Genome Centre
Department of Genetics
855 California Avenue
Palo Alto CA 94304
USA

☎
[redacted]

Dr Maynard Ilson
University of Washington
Molecular Biotechnology GJ-10
Mason Road
Fluke Hall
Seattle WA98195
USA

☎
[redacted]

Professor Yoshiyuki Sakaki
Human Genome Centre
Institute of Medical Science
The University of Tokyo
4-6-1 Shirokamedai Mimato-ku
Tokyo 108
Japan

☎
[redacted]

Dr Melvin I. Simon
California Institute of Technology
Division of Biology 147-75
Pasadena
California 91125
USA

☎
[redacted]

Dr James D. Watson
Cold Spring Harbor Laboratory
One Bungtown Road
P.O.Box 100
Cold Spring Harbor
New York 11724
USA

☎
[redacted]

Dr Richard Wilson
Washington University Medical School
Genome Sequencing Centre
4444 Forest Park
Box 8501
St Louis MO 63108
USA

☎
Fax [redacted]

2nd NHGRI Large-Scale Sequencing QA Exercise

METHOD

- Selected four finished clones, at random, totaling 200 kb, from each participating sequencing group (all NHGRI human plus *D. melanogaster*)
- Data checked was selected from that deposited as 'finished' as of September, 1997
- Assigned each set of four clones to two checkers chosen from among the participants; groups exchanged data files and bacterial isolates/DNA
- Checkers re-assembled files and analyzed data. If error rate was better than 1 in 2000, resolved discrepancies by further analysis (resequencing).
- Each group responded to checker's reports
- Most groups checked assembly by restriction analysis

Total number of clones available for checking *as of 9/97*: 420
Total number of clones selected for exercise: 37 (a total of 1.7 Mb tested)

2nd NHGRI Large-Scale Sequencing QA Exercise

RESULTS

Single-base discrepancies–number of clones at[a]:

<1/10000	1/10000-1/5000	1/5000-1/2000	>1/2000	Total
22	10[b]	1	3	36*

[a] These numbers are based on the results that indicated the higher error rate among the two reports, for each individual clone; these numbers do not take into account the producer's responses.

[b] For 7 out of the 10 clones in this category, one of the two checkers actually evaluated those clones as having fewer than 1 in 10000 errors.

- Total number of single-base discrepancies (conservative aggregate of two checkers): 230/1.7 Mb. Total excluding the clones worse than 1 in 2000: 120/1.59 Mb

- About 2/3rds (133) of the single-base discrepancies were substitutions, 1/3rd (73) were insertions or deletions, based on 206 cases of single-base errors where precise information was provided

Other errors (not exclusive of single-base errors)

1 mis-assembly, origin unknown
1 possible mis-assembly (1900-base deletion); may be a clone instability
1 clear clone instability (~250 bp deletion)
1 likely clone instability (~650 bp deletion)
1 annotated gap closed (75 bp)
*1 wrong clone sent (clone tracking error)

2nd NHGRI Large-Scale Sequencing QA Exercise

SUMMARY

- **Caveats:** Variability due to sampling; variability in checking

- **Most groups are sequencing at or very close to standards:** Most groups are at 1 in 10000 or better, summed over all clones. Numbers in the table are conservative and do not include the producer's responses, consideration of which will improve the error rates. However, most of the producers responses agree with the checkers' reports.

- **Good concordance between checkers' reports**: For single-base errors, both checkers agreed on the general quality of the project (according to the bins in Table 1) 28 of 37 times, and were very close in all other cases. In 11 of 19 clones where error type and location appear in the report, there is at least a 50% overlap in the precise identified errors. But there were still some puzzling differences between the identified errors in an individual clone, especially when there were a lot of errors or trace data were considered poor by checker. For other types of error (deletions, etc.), both checkers agreed in all but one case.

- **The exercise reveals useful information about the kinds of error:** Clone instabilities (small deletions) were a small but significant problem—small deletions may be hard to detect with routine protocols. Single-base errors often occur in regions where sequence data quality is good—more than half could be resolved unambiguously by re-editing the original data without need to re-sequence (36/53 errors). (Some of this was confirmed by resequencing).

The National Human Genome Research Institute

NHGRI Policy on Release of Human Genomic Sequence Data

March 7, 1997

At the Second International Strategy Meeting on Human Genome Sequencing (Bermuda, 1997), attendees affirmed the principle that was set out at the First (1996) International Strategy meeting, that primary genomic sequence should be rapidly released. Specifically, the report of the first meeting stated that "sequence assemblies should be released as soon as possible; in some centres, assemblies of greater than 1 kb would be released automatically on a daily basis." The discussions at the 1997 meeting confirmed NHGRI's conclusions that it is extremely important for its large-scale sequencing program to be functioning in a manner consistent with this principle, that such rapid release is technically feasible, and that such unfinished DNA sequence data have already been found to be useful by the larger scientific community. NHGRI has determined, therefore, that its grantees engaged in large-scale genomic DNA sequencing should now be automatically releasing sequence assemblies of 2 kb or larger within 24 hours of their generation. (the trigger for data release is 2 kb, instead of 1 kb, in order to ensure that the released sequence be comprised of at least two sequence reads. Investigators who wish to release smaller assemblies may do so.) Any laboratory funded by NHGRI for large-scale human genomic sequencing must develop and submit to NHGRI a plan to implement such a data release program, which must be implemented within one month of its being approved by NHGRI. No non-competing or competing renewal will be funded until an acceptable plan has been approved. Mandatory data release as described above will be made a condition of the award for any grant funded by NHGRI for large-scale human sequencing.

Genomic & Genetic Data
Grant Information
Policy & Public Affairs
Offsite Resources
Human Genome Project
About NHGRI
Intramural Research
Search

webmaster@nhgri.nih.gov

Genomic and Genetic Data | Grant Information | Policy and Public Affairs |
Offsite Resources | The Human Genome Project | About NHGRI | Intramural Research | Keyword Search

webmaster@nhgri.nih.gov

The National Human Genome Research Institute

NIH-DOE Guidelines for Access to Mapping and Sequencing Data and Material Resources

The information and resources generated by the Human Genome Project have become substantial, and the interest in having access to them is widespread. It is therefore desirable to have a statement of philosophy concerning the sharing of these resources that can guide investigators who generate the resources as well a those who wish to use them.

A key issue for the Human Genome Project is how to promote and encourage the rapid sharing of material and data that are produced, especially information that has not yet been published or may never be published in its entirety. Such sharing is essential for progress toward the goals of the program and to avoid unnecessary duplication. It is also desirable to make the fruits of genome research available to the scientific community as a whole as soon as possible to expedite research in other areas.

Although it is the policy of the Human Genome Project to maximize outreach to the scientific community, it is also necessary to give investigators time to verify the accuracy of their data and to gain some scientific advantage from the effort they have invested. Furthermore, in order to assure that novel ideas and inventions are rapidly developed to the benefit of the public, intellectual property protection may be needed for some of the data and materials.

After extensive discussion with the community of genome researchers, the advisors of the NIH and DOE genome programs have determined that consensus is developing around the concept that a 6 month period from the time data or materials are generated to the time they are made available publicly is a reasonable maximum in almost all cases. More rapid sharing is encouraged.

Whenever possible, data should be deposited in public databases and materials in public repositories. Where appropriate repositories do no exist or are unable to accept the data or materials, investigators should accommodate requests to the extent possible.

The NIH and DOE genome programs have decided to require all applicants expecting to generate significant amounts of genome data and materials to describe in their application how and when they plan to make such data and materials available to the community. Grant solicitations will specify this requirement. These plans in each application will be reviewed in the course of peer review and by staff to assure they are reasonable and in conformity with program philosophy. If grant is made, the applicant's sharing plans will become a condition of the award and compliance will be reviewed before continuation is provided. Progress reports will be asked to address the issue.

Genomic & Genetic Data
Human Genome Project
Grant Information
About NHGRI
Policy & Public Affairs
Intramural Research
Offsite Resources
Search

webmaster@nhgri.nih.gov

Professor Nobyoshi Shimizu
Department of Molecular Biology
Keio University School of Medicine
35 Shinanomachi
Shinjuku-ku
Tokyo 160-8582
Japan
[redacted]

Dr Douglas Smith
Director, GTC Sequencing Center and Technology Development
Genome Therapeutics Corporation
100 Beaver Street
Waltham MA 02453
USA
[redacted]

Dr Jean Weissenbach
GENOSCOPE
Centre national de sequencage
2 rue Gaston Cremieux
CP 5706
91057 Evry Cedex
France
[redacted]

Dr. Huanming Yang
Professor and Director, Genome Center
Institute of Genetics
Chinese Academy of Sciences
Datun Road, Beijing, 100101
China
[redacted]

Mack

Distibution:

Mailed to:

David	Cox	W/O form
Ronald	Davis	
Richard	Gibbs	
Leroy	Hood	
Eric	Lander	
Maynard	Olson	
Greg	Schuler	W/O form
Douglas	Smith	
Robert	Waterston	

BCC:
Marvin Frazier
Michael Morgan
Elke Jordan

DEPARTMENT OF HEALTH & HUMAN SERVICES

Public Health Service

National Institutes of Health
National Human Genome
Research Institute
31 Center Drive, MSC 2033
Building 31, Room B2B-07
Bethesda, MD 20892-6050
(301) 496-7531
FAX (301) 480-2770

«Title» «FirstName» «LastName»
«Company»
«Address1»
«Address2»
«City», «State» «PostalCode»

Dear «Title» «LastName»:

On behalf of the National Institutes of Health, the U.S. Department of Energy and the Wellcome Trust, I am writing to announce the scheduling of the Sixth International Strategy Meeting on Human Genome Sequencing. The meeting will be held on January 13, 2000 at the Marriott Hotel in Walnut Creek, California, near the DOE's Joint Genome Institute. The purpose of the meeting will be to review progress of the international consortium of laboratories that are sequencing the human genome, to discuss critical issues related to completing the working draft sequence of the human genome by Spring, 2000 and to plan for the ultimate production of the finished sequence.

As you will recall, it has been agreed that rapid submission of all sequence data to the public databases in accord with the Bermuda agreements is a condition for continued attendance at the International Strategy meetings. The National Center for Biotechnology Information (NCBI) summarizes database submissions on a weekly basis; this report can be found at http://www.ncbi.nlm.nih.gov/genome/seq/weekly_report.html. NHGRI staff review this report regularly to monitor laboratories' compliance with the international agreements, and you are encouraged to check it regularly to ensure that data you have deposited is being correctly logged to your center.

In determining the final invitation list for the January meeting, we will be reviewing data production and data submission from all preliminary invitees during early December. To assist us in this review, we request that you send a report of your actual production during the period September through November 1999 to Jane Peterson and Mark Guyer (jane_peterson@nih.gov; mark_guyer@nih.gov) by December 10, 1999. Please report the number of reads attempted and the number of successful reads (meaning reads that produce data actually used in assemblies), as well as the amount of data that was deposited in the public databases during this period and the percent of working draft sequence. The data should be provided in the format attached. An electronic version will be e-mailed to you. We expect that the database deposits should approximate the amount of production, as reported in reads. If there is a significant discrepancy, it would be very helpful if you would explain the basis of the difference. Furthermore, at the last International Meeting, the participants estimated the amount of sequence they expected their laboratory to produce from September through November 1999. If there is a significant difference between the amount of production you estimated at that time, and the amount your laboratory actually produced, it would be very helpful if you would explain the reasons for that difference as well. Please also make read projections for next quarter (12/1/99 to 2/29/00) and make changes to the yearly total estimates in the attached form.

A reception with a cash bar will be held the evening before the at the hotel. More information about the reception and the agenda for the meeting will be sent to you once your production report and information about data deposition has been received.

Sincerely yours,

Francis S. Collins, M.D., Ph.D.
Director

Enc: 2

As per Dr. Collins' letter, please complete the following chart for your sequencing center, by filling in the information in the shaded areas. For reference, the chart containing equivalent data from last quarter's meeting is included. Please return to Jane Peterson and Mark Guyer (jane_peterson@nih.gov; mark_guyer@nih.gov) by December 10, 1999. If you would prefer, you may complete this electronically – the form will be e-mailed to you shortly.

Sequencing Production Figures (9/1/99 to 11/30/99)
Sixth International Strategy Meeting on Human Genome Sequencing

Center	Regions	Size (Mb)	Yearly (4/99 -3/00) Projected Reads (k)		Current Quarter 9/1/99 - 11/30/99					Next Quarter 12/1/99 - 2/29/99
			Sept 99 Estimates	"New" Estimates	Projected Reads (k)	**Attempted Reads (k)**	**Successful Reads (k)**	**Genbank (kb)**	**% Working Draft**	**Projected Reads (k)**
Baylor	3, 12, X	230	3,100		660					
GBF	9 , 21	6	300		50					
Genoscope	14	85	1,400		300					
IMB	8,21,X	50	1,500		180					
JGI	5,16,19	250	6,400		2,100					
Keio	2,6,8,21,22	23	230		110					
MPIMG	17,21,X	6.9	300		40					
Riken	11q, 18, 21	160	2,100		520					
Sanger Centre	1,6,9,10,13, 20,22, X	1000	12,000		4,200					
Wash U	2,3,4,7,8,11, 15,17,18,Y	900	7,900		2,300					
WIBR	2,3,4,7,8,11, 15,17,18,Y	900	8,000		2,900					
Beijing	3p	30	500		100					
U. Wash (Hood)	14,15	50	170		40					
U. Wash (Olson)										
Stanford	8		290		137					
GTC	10	50	450		150					
Total		**2671**	**44,640**		**13,787**					

Sequencing Production Figures (6/1/99 to 8/31/99)
Fifth International Strategy Meeting on Human Genome Sequencing

Center	Regions	Size	Yearly (4/99 -3/00) Projected Reads (k)		Current Quarter 6/1/99 – 8/31/99				Next Quarter 9/1/99 - 11/30/99
		(Mb)	May 99 Estimates	Sept 99 Estimates	Projected Reads (k)	%WD	Actual Reads (k)	Genbank (kb)	Projected Reads (k)
Baylor	3, 12, X	230	3,000	3,100	420	75	462	261	660
GBF	9 , 21	6	6,000	0,300					50
Genoscope	14	85	1,400	1,400	200	60	100	118	300
IMB	8,21,X	50	2,100	1,500	375	50	180	32	180
JGI	5,16,19	250	4,500	6,400	860	90	687	461	2,100
Keio	2,6,8,21,22	23	300	230	70	60	11	75	110
MPIMG	17,21,X	6.9	300	300	50	40	40	12	40
Riken	11q, 18, 21	160	1,900	2,100	360	50	136	195	520
Sanger Centre	1,6,9,10,13, 20,22, X	1,000	12,000	12,000	3,000	33	1,300	941	4,200
Wash U	2,3,4,7,8,11, 15,17,18,Y	900	7,900	7,900	977		865	559	2,300
WIBR	2,3,4,7,8,11, 15,17,18,Y	900	8,000	8,000	1,230	90	837	296	2,900
Beijing	3p	30		500			12.5		100
U. Wash (Hood)	14,15	50		170		90	27		40
U. Wash (Olson)									
Stanford	8			290		90			137
GTC	10	50		450		90	5		150
Total		**2,671**	**41,400**	**44,640**	**7,542**	**50**	**4,652**	**2,950**	**13,787**

Wetterstrand, Kris (NHGRI)

From: Peterson, Jane (NHGRI)
Sent: Tuesday, November 16, 1999 1:57 PM
To: 'David Cox'; Davis, Ronald; 'Richard Gibbs'; 'Lee Hood'; 'Eric Lander'; 'Maynard Olson'; Schuler, Greg (NLM); 'Doug Smith'; 'Bob Waterston'
Cc: [redacted]
Whittington, Peggy (NHGRI); Wetterstrand, Kris (NHGRI)
Subject: Date for International Meeting January 12, 2000

Importance: High

Last Friday we sent you an e-mail followed by a letter informing you of the date in January, 2000 for the next International Strategy Meeting to be held near the JGI in Walnut Creek, California. Unfortunately the date of the meeting given in that correspondence was incorrect. The **correct date is January 12, 2000**. Please make a note of this. For your information, we are planning a reception on the night of the January 11, 2000 for all of the attendees. More information will follow.

Jane L. Peterson, Ph.D.
Program Director, Large Scale Sequencing
National Human Genome Research Institute
Building 38A, Room 614
38 Library Drive MSC 6050
Bethesda, MD 20892-6050
[redacted]

Dr David Bentley
Wellcome Trust Genome Campus
The Sanger Centre
Hinxton
Cambridgeshire CB10 1SA
UK
☎ 01223 832244
Fax: 01223 494919
Email: drb@sanger.ac.uk

Dr Helmut Blöcker
Head of Department
Genome Analysis
GBF
Mascheroder Weg 1
D-38124 Braunschweig
Germany
☎ 00 49 531 6181 220
Fax: 00 49 531 6181 292
Email: bloecker@gbf.de

Dr Elbert Branscomb
Department of Energy Joint Genome
Institute
2800 Mitchell Drive
Walnut Creek CA 94598
USA
☎ 00 1 925 296 5608
Fax: 00 1 925 296 5710
Email: branscomb1@llnl.gov

Mr Graham Cameron
The European Bioinformatics Institute
The Wellcome Trust Genome Campus
Hinxton
Cambridgeshire CB10 1SD
U.K
☎ 01223 494467
Fax: 01223 494470
Email: cameron@ebi.ac.uk

Asif Chinwalla
Director of Informatics
Genome Sequencing Center
Washington University School of
Medicine
4444 Forest Blvd.
St Louis MO 63108
USA
☎ 00 1 314 286 1811
Fax: 00 1 314 286 1810
Email: achinwal@watson.wustl.edu

Dr Francis Collins
Director
National Human Genome Research
Institute
31 Center Drive MSC 2152
Building 31 Room 4B09
Bethesda MD 20892-2152
USA
☎ 00 1 301 496 0844
Fax: 00 1 301 402 0837
Email: fc23@nih.gov

Dr David Cox
Professor of Genetics and Paediatrics
Stanford University School of Medicine
300 Pasteur Drive
Stanford CA 94305-5120
USA
☎ 00 1 650 725 8043
Fax: 00 1 650 725 8058
Email: sbasso@shgc.stanford.edu

Ian Dunham
The Sanger Centre
The Wellcome Trust Genome Campus
Hinxton
Cambridgeshire CB10 1RQ
UK
☎ 01223 494948
Fax: 01223 494919
Email: id1@sanger.ac.uk

Dr Nancy Federspiel
Associate Director
Stanford DNA Sequencing and
Technology Center
855 California Avenue
Palo Alto CA 94304
USA
☎ 00 1 650 812 1971
Fax: 00 1 650 812 1975
Email: nfeder@sequence.Stanford.edu

Dr Marvin Frazier
U.S. Department of Energy
Health, Effects & Life Sciences Research
ER-72 GTN
19901 Germantown Road
Germantown MD 20874-1290
USA
☎ 00 1 301 903 5468
Fax: 00 1 301 903 8521
Email: Marvin.Frazier@science.doe.gov

Dr Asao Fujiyama
Team Leader, Genome Sciences Center at RIKEN
National Institute of Genetics
Division of Human Genetics
1111 Yata
Mishima
Shizuoka 411-8540
Japan
☎ 00 81 559 81 6796
Fax: 00 81 559 81 6716
Email: afujiyam@lab.nig.ac.jp

Dr Richard Gibbs
Department of Molecular & Human Genetics
Baylor College of Medicine
One Baylor Plaza
BCM-226
Houston TX 77030
USA
☎ 00 1 713 798 6589
Fax: 00 1 713 798 5741
Email: agibbs@bcm.tmc.edu

Dr Trevor Hawkins
Department of Energy Joint Genome Institute
2800 Mitchell Drive
Walnut Creek CA 94598
USA
☎ 00 1 925 296 5682
Fax: 00 1 925 296 5710
Email: tlhawkins@lbl.gov

Dr Tim Hubbard
The Sanger Centre
The Wellcome Trust Genome Campus
Hinxton
Cambridgeshire CB10 1RQ
UK
☎ 01223 494983
Fax: 01223 494919
Email: th@sanger.ac.uk

Dr Lauren Linton
Co-Director, Sequencing
Whitehead Institute/MIT Center for Genome Research
320 Charles Street
Cambridge MA 02141
USA
☎ 00 1 617 258 0900
Fax: 00 1 617 258 0903
Email: lml@genome.wi.mit.edu

Dr Richard McCombie
Associate Professor
Cold Spring Harbor Laboratory
1 Bungtown Road
Cold Spring Harbor NY 11724
USA
☎ 00 1 516 367 8884
Fax: 00 1 516 367 8874
Email: mccombie@cshl.org

Professor John McPherson
Genome Sequencing Center
Washington University School of Medicine
4444 Forest Park Blvd
St Louis MO 63108
USA
☎ 00 1 314 286 1848
Fax: 00 1 314 286 1810
Email: jmcphers@watson.wustl.edu

Dr Anup Madan
Research Scientist
University of Washington
Department of Molecular Biotechnology
Box 357730
1705 NE Pacific
Room K354 HSB
Seattle, Washington 98195-7730
[redacted]

Dr Elaine Mardis
Genome Sequencing Center
Washington University School of
Medicine
4444 Forest Blvd.
St Louis MO 63108
USA
[redacted]
mail: emardis@watson.wustl.edu

Dr Shinsei Minoshima
Center for Genomic Medicine
Keio University School of Medicine
35 Shinanomachi, Shinjuku-ku
Tokyo 160-8582
Japan
☎ [redacted]
Fax [redacted]

Dr [redacted]
The [redacted]
183 Euston Road
London NW1 2BE
UK
[redacted]
Email:
[redacted]

Dr Jane Peterson
Program Director
Large Scale Sequencing
National Institutes of Health
National Center for Human Genome
Research
Building 38A Room 614
Bethesda MD 20892
[redacted]
Fax: 00 1 301 480 2770
Email: pet[redacted]

Dr Debbie Poole
Programme Manager
The Wellcome Trust Genome Campus
Hinxton Hall
Hinxton
Cambridgeshire CB10 1RQ
[redacted]
Fax: [redacted] 495103
Email: d.poole@hinxton.wellcome.ac.uk

Dr Julianne Ramser
Max-Planck-Institut fuer Molekulare
Genetik
Ihnestr. 73
D-14195 Berlin
Germany
[redacted].de

The Wellcome Trust
183 Euston Road
[redacted]
UK
☎ [redacted]
[redacted]
Email: s[redacted]k

Dr Jane Rogers
The Sanger Centre
The Wellcome Trust Genome Campus
Hinxton
Cambridgeshire CB10 1RQ
UK
☎ [redacted]

Professor André Rosenthal
Institute of Molecular Biotechnology
Department of Genome Analysis
Beutenbergstrasse 11
07745 Jena
Germany
☎ [redacted]

Professor Yoshiyuki Sakaki
Project Leader, Genome Sciences Center
at RIKEN
Human Genome Center
Institute of Medical Science
University of Tokyo
4-6-1 Shirokanedai
Minato-ku
Tokyo 108-8639
Japan
☎ [redacted]
Fax: 00 81 3 5449 5445
Email: sakaki@ims.u-tokyo.ac.jp

Dr Takashi Sasaki
Center for Genomic Medicine
Keio University School of Medicine
35 Shinanomachi
[redacted]

Professor Nobyoshi Shimizu
Department of Molecular Biology
Keio University School of Medicine
35 Shinanomachi
Shinjuku-ku
Tokyo 160-8582
Japan
☎ [redacted]

Dr Douglas Smith
Director, GTC Sequencing Center and
Technology Development
Genome Therapeutics Corporation
100 Beaver Street
Waltham MA 02453
USA
[redacted]
Fax: [redacted]

Dr Jean Weissenbach
GENOSCOPE
Centre national de sequencage
2 rue Gaston Cremieux
CP 5706
[redacted]

Dr. Huanming Yang
Professor and Director, Genome Center
Institute of Genetics
Chinese Academy of Sciences
Datun Road, Beijing, 100101
China
☎ [redacted]

Dr David Bentley
Wellcome Trust Genome Campus
The Sanger Centre
Hinxton
Cambridgeshire CB10 1SA
UK
[redacted]

Dr Helmut Blöcker
Head of Department
Genome Analysis
GBF
Mascheroder Weg 1
D-38124 Braunschweig
Germany
[redacted]

Dr Elbert Branscomb
Department of Energy Joint Genome Institute
2800 Mitchell Drive
Walnut Creek CA 94598
USA
00 [redacted]

Mr Graham Cameron
The European Bioinformatics Institute
The Wellcome Trust Genome Campus
Hinxton
Cambridgeshire CB10 1SD
U.K
[redacted]

Asif Chinwalla
Director of Informatics
Genome Sequencing Center
Washington University School of Medicine
4444 Forest Blvd.
St Louis MO 63108
USA
([redacted]

Dr Francis Collins
Director
National Human Genome Research Institute

31 Center Drive MSC 2152
Building 31 Room 4B09
Bethesda MD 20892-2152
USA
[redacted]

Dr David Cox
Professor of Genetics and Paediatrics
Stanford University School of Medicine
300 Pasteur Drive
Stanford CA 94305-5120
USA
[redacted]

Ian Dunham
The Sanger Centre
The Wellcome Trust Genome Campus
Hinxton
Cambridgeshire CB10 1RQ
UK
[redacted]

Dr Nancy Federspiel
Associate Director
Stanford DNA Sequencing and Technology Center
855 California Avenue
Palo Alto CA 94304
USA
[redacted]

Dr Marvin Frazier
U.S. Department of Energy
Health, Effects & Life Sciences Research
ER-72 GTN
19901 Germantown Road
Germantown MD 20874-1290
USA
[redacted]

Dr Asao Fujiyama
Team Leader, Genome Sciences Center at RIKEN
National Institute of Genetics
Division of Human Genetics
1111 Yata

Mishima
Shizuoka 411-8540
Japan
[redacted]

Dr Richard Gibbs
Department of Molecular & Human Genetics
Baylor College of Medicine
One Baylor Plaza
BCM-226
Houston TX 77030
USA
[redacted]

Dr Trevor Hawkins
Department of Energy Joint Genome Institute
2800 Mitchell Drive
Walnut Creek CA 94598
USA
[redacted]

Dr Tim Hubbard
The Sanger Centre
The Wellcome Trust Genome Campus
Hinxton
Cambridgeshire CB10 1RQ
UK
[redacted]

Dr Lauren Linton
Co-Director, Sequencing
Whitehead Institute/MIT Center for Genome Research
320 Charles Street
Cambridge MA 02141
USA
[redacted]

Dr Richard McCombie

Associate Professor
Cold Spring Harbor Laboratory
1 Bungtown Road
Cold Spring Harbor NY 11724
USA
[redacted]

Professor John McPherson
Genome Sequencing Center
Washington University School of Medicine
4444 Forest Park Blvd
St Louis MO 63108
USA
[redacted]

Dr Anup Madan
Research Scientist
University of Washington
Department of Molecular Biotechnology
Box 357730
1705 NE Pacific
Room K354 HSB
Seattle, Washington 98195-7730
[redacted]

Dr Elaine Mardis
Genome Sequencing Center
Washington University School of Medicine
4444 Forest Blvd.
St Louis MO 63108
USA
[redacted]

Dr Shinsei Minoshima
Center for Genomic Medicine
Keio University School of Medicine
35 Shinanomachi, Shinjuku-ku
Tokyo 160-8582
Japan
[redacted]

Dr Michael Morgan
The Wellcome Trust
183 Euston Road
London NW1 2BE

UK
[redacted]

_
Dr Jane Peterson
Program Director
Large Scale Sequencing
National Institutes of Health
National Center for Human Genome Research
Building 38A Room 614
Bethesda MD 20892
USA
[redacted]

Dr Debbie Poole
Programme Manager
The Wellcome Trust Genome Campus
Hinxton Hall_
Hinxton
Cambridgeshire CB10 1RQ
UK
[redacted]

Dr Julianne Ramser
Max-Planck-Institut fuer Molekulare Genetik
Ihnestr. 73
D-14195 Berlin
Germany
[redacted]

Dr Siân Renfrey
The Wellcome Trust
183 Euston Road
London NW1 2BE
UK
[redacted]

Dr Jane Rogers
The Sanger Centre
The Wellcome Trust Genome Campus
Hinxton
Cambridgeshire CB10 1RQ
UK

[redacted]

Professor André Rosenthal
Institute of Molecular Biotechnology
Department of Genome Analysis
Beutenbergstrasse 11
07745 Jena
Germany

[redacted]

Professor Yoshiyuki Sakaki
Project Leader, Genome Sciences Center at RIKEN
Human Genome Center
Institute of Medical Science
University of Tokyo
4-6-1 Shirokanedai
Minato-ku
Tokyo 108-8639

[redacted]

Dr Takashi Sasaki
Center for Genomic Medicine
Keio University School of Medicine
35 Shinanomachi
Shinjuku-ku
Tokyo 160-8582

[redacted]

Dr. Jeffery A. Schloss
Program Director
Technology Development Co-ordination
National Human Genome Research Institute
National Institutes of Health
Bldg 38A, Room 614
38 Library Drive, MSC 6050
Bethesda, MD 20892-6050

[redacted]

FYI

RESOURCES AVAILABLE

Name of participant : Mark Adams / Craig Venter - TIGR

Nature of resources available (software, maps, clones etc.)

Software

TIGR Assembler - sequence assembly

HBQCM - hexamer-based composition tool

yank - GenBank extraction software

TIGR's sybase schema

Human cDNA Database - >355,000 ESTs, >51,000 THC assemblies

Available via:

e-mail request to arkerlav@tigr.org (Tony Kerlavage)
cDNA clones through TIGR/ATCC and www

Any conditions attached:

None.

RESOURCES AVAILABLE

Name of participant : ANSORGE/EMBC

Nature of resources available (software, maps, clones etc.)

- SEQUENCING TECHNOLOGY - 100kb/per run

GENESKIPPER - ASSEMBLY PROGRAMM
+ SEQUENCE ANALYSIS

RAN-DI (Random-Direct) strategy -
assembling first 80-100 clones
randomly sequenced
+ all ECOR1 fragments, finish with DIRECT strategy
→ NO CLONING GAPS OBSERVED.

Available via:

FAX or e-mail
to EMBC

Any conditions attached:

RESOURCES AVAILABLE

Name of participant :

Tony Carrano
Lawrence Livermore National Laboratory

Nature of resources available (software, maps, clones etc.)

Resource	Availability
High-resolution, metric map of chromosome 19	Published version available in Dec issue of Nature Genetics. Detailed version available by collaboration
Arrayed cosmid libraries of human chromosomes	Through major genome centers. Soon to be available through the UK and German resource centers.
IMAGE collection of cDNAs	Available through industry and resource centers.
DNA sequence sample tracking software	Contact Tom Slezak @ LLNL
Clone fingerprinting assembly and database software	Contact Tom Slezak @ LLNL
Mapping infrastructure resource (creating high-resolution sequence ready maps in cosmids and BACs)	Contact Tony Carrano @ LLNL

Available via:

see above

Any conditions attached:

Creating maps as part of the mapping infrastructure resource would require funding.

RESOURCES AVAILABLE

Name of participant : RICHARD DURBIN

Nature of resources available (software, maps, clones etc.)

SOFTWARE; ACEDB database system

SAM (Cari Soderlund) marker assembly/edit/viewer

FPC (" ") fingerprint " " " SOON!

AUTOEDIT - sequence automatic editor for assemblies (Richard Mott)

MSPCRUNCH/BELVU/DOTTER - sequence analysis/viewing tools

Available via: ANONYMOUS FTP (FTP.SANGER.AC.UK)

email: RD@SANGER.AC.UK

Any conditions attached:

NO COMMERCIALISATION (use by companies OK)

Resources Available

Name of participant:

Glen A. Evans

Nature of Resources:

1. Chromosome 11 Sequencing DataBases

   YAC/STS coordinates database
   cosmid end sequence database
   YAC-cosmid coordinate database
   Primers (new STSs)
   Homology/Identities listed by match significance

   Chromosome 11 sequencing data (complete cosmid/PAC sequences)
   11p15 project, 11p12 project
   WWW http://mcdermott.swmed.edu/
   Genbank

2. Clone libraries

   chromosome 11 cosmid 5X, arrayed
   chromosome 11 YAC ?X, arrayed (T. Shows/N. Nowak, RP)
   chromosome 11 and 15 PAC set in preparation

   (can be made available on request to G. Evans)

3. Software

   Mermade driver software for 192 channel synthesizer
   Primer prediction software for primer directed walking
   SUMU Lab sample tracking software
   Robotics control software for Biomek
   Data Inspector software for sequence quality control

   WWW http://mcdermott.swmed.edu/

4. Hardware specifications and construction plans

   Prepper III miniprep robot
   Mermade 192 channel oligonucleotidesynthesizer
   Lab workstations
   TREC multigel controller

Lab workstation plans and ordering information

WWW http://mcdermott.swmed.edu/

Available via:

WWW http://mcdermott.swmed.edu/

Any conditions attached:

Data resources are made available within 6 months after generation.

Hardware and software are supplied without warranty and without support other than helpful hints when needed. Hardware specifications and plans are available to all non-commerical users.

RESOURCES AVAILABLE

Name of participant : Chris Fields

Nature of resources available (software, maps, clones etc.)

Chris Fields

GSDB (complete, genome-scale relational DB) scheduled for operational mid-summer

GSDB "Annotator" multiplatform client interface (view/edit) available free mid summer.

Available via: http://www.ncgr.org

Any conditions attached: none

RESOURCES AVAILABLE

Name of participant : Richard A. Gibbs

Nature of resources available (software, maps, clones etc.)

- X chromosome mapped reagents - including binned cosmids (≥ 2,000) -
- Sequences, cosmids and the shotgun libraries from >1mb of human DNA from X, ch12 + ch17 available,
- matched cosmid/cDNA pairs available from X-chromosome, from C.C. Lee.

Available via:

All X chromosome + ch12 resources are described in their respective web pages.

Any conditions attached:

NO

RESOURCES AVAILABLE

Name of participant : Trevor Hawkins

Nature of resources available (software, maps, clones etc.)

>15,000 Human mapped STSs

>6,500 Mouse mapped SSRs

GRACE/BASS Gel analysis and basecalling software, UNIX based.

Primer Picking software (PRIMER 2.2)

LabBase database system

Available via: http://www-genome.wi.mit.edu

Any conditions attached: None.

RESOURCES AVAILABLE

Name of participant : LaDeana Hillier

Nature of resources available (software, maps, clones etc.)

SOFTWARE :

GETLANES (tracking gel images)

RETRAK (UNIX interface for editing lane tracking)

TPP (trace processing software)

PHRED (base calling)

PHRAP (sequence assembly)

FINISH (following shotgun completion, finish selects reads to contiguate & improve sequence quality)

DACE (implementation of a laboratory notebook tracking system in ACEDB)

other software tools are also available

Available via: HTTP:// genome.wustl.edu/gschmpg.html

PHRED & PHRAP available: phg@u.washington.edu

ACEDB code available: ncbi.nlm.nih.gov /pub/repository/acedb

Any conditions attached:

retrak & tpp are still under intensive development.

RESOURCES AVAILABLE

Name of participant : PIETER DE JONG

Nature of resources available (software, maps, clones etc.)

Human PAC library 120 kb average insert (16-fold redundant, ~1200 384 well dishes)
(male donor, DNA from blood)

Human PAC library (~5-fold redundant) not yet arrayed; 150 kb insert.
(female donor, DNA from blood

Human BAC library : in progress,
expect to deliver 10-fold redundant
by May '96 and 20-fold by Summer '96.

Available via: PdJ, Roswell Park Cancer Institute

Any conditions attached:

- No secondary distribution of library, no problems to distribute individual clones (no ties attached).
- Cost-recovery of labor/plasticware/mailing costs for library replicates.

RESOURCES AVAILABLE

Name of participant: Dr. Hans Lehrach

Nature of resources available (software, maps, clones etc.)

The Resource Centre distributes high-density gridded filters of genomic libraries, cultures of individual library clones, or (in the future) PCR pools.

The table below gives details of those genomic libraries for which this service is now available, in the near future this will be supplemented with libraries from the I.M.A.G.E. consortium:

Library name	Description	Number
Cosmid (Human)		
L4/FS1	Chromosome 1 specific cosmid library	112
L4/FS6	Chromosome 6 specific cosmid library	109
L4/FS7	Chromosome 7 specific cosmid library	113
L4/FS11	Chromosome 11 specific cosmid library	107
L4/FS13	Chromosome 13 specific cosmid library	108
L4/FS17	Chromosome 17 specific cosmid library	105
L4/FS18	Chromosome 18 specific cosmid library	111
L4/FS21	Chromosome 21 specific cosmid library	102
L4/FS22	Chromosome 22 specific cosmid library	106
L4/FSC X/LA	Chromosome X specific cosmid library	101
L4/FSC X	Chromosome X specific cosmid library	104
Cosmid (other)		
L4/S.Pombe	S.pombe specific cosmid library	60
L4/B/S.Pombe	S.pombe specific cosmid library	61
Fugu-Cosmid	Fugu DNA partial cut with MboI in Lawrist4 and DH10B	66
P1		
P1 Human	Total Genomic P1 Human Library	700
MP1 Mouse P1 library	Total Genomic Mouse C57/Black6 P1 Library	703
pomP1	Schizosaccharomyces pombe (wt 972 h-) P1 library	705
PAC		
Human PAC	Human PAC library brought by Peter de Jong	704

RESOURCES AVAILABLE (Continued)

Name of participant: Dr. Hans Lehrach

Nature of resources available (software, maps, clones etc.)

(Continued from previous page)

Library name	Description	Number
YAC (Human)		
4X YAC	Human YAC library	900
4Y YAC	Human YAC library	901
CEPH YAC	Human CEPH YAC library	904
LSXY	Human YAC library	912
C3H YAC	Mouse YAC library	902
YAC (other)		
St.Marys Mouse YAC RAD52	Mouse YAC library from female C57BL/10 in host strain which is recombination deficient due to mutation in RAD52	909
C57 YAC	Mouse YAC library	903
Whitehead Mousc YAC I	Large insert Mouse YAC library constructed at the Whitehead Institute for Biomedical Research/MIT Center for Genome Research	910
pomYAC	Schizosaccharomyces pombe (wt 972 h-) YAC library	913
ICRF Pig YAC	Pig YAC library	907
LMUB Pig YAC	Pig YAC library from Lymphocytes (~300KB average inserts) from Ludwig Maximillian Univ.Muenchen	911
cDNA (Human)		
Human fetal brain cDNA	Human foetal brain cDNA made from 17 week embryo polyA+RNA	507
HFL cDNA	cDNA using dT primed polyA+ purified RNA from 21 weeks old human fetal liver	512
HTE cDNA	cDNA using dT primed polyA+ purified RNA from 21 weeks old human fetal thymus	508
HPO cDNA	cDNA from 21 weeks human foetal lung, poly dT primed, directionally cloned, excise enzyme MluI	515
cDNA (other)		
MBR cDNA	Mouse adult brain cDNA,synth: oligo dT primed,directionally cloned; cloning site: NotI/SalI; 1.5kb average insert size	510

RESOURCES AVAILABLE (continued)

Name of participant: Dr. Hans Lehrach

Nature of resources available (software, maps, clones etc.)

(see previous pages)

Available via:

The Resource Centre/Primary Database of the German Human Genome Project,
Max-Planck-Institut für Molekulare Genetik,
(Abteilung Lehrach),
Ihnestraße 73,
14195 Berlin (Dahlem)
GERMANY

Tel: +49 [redacted]

Fax: [redacted]

WWW: http://rzpd.rz-berlin.mpg.de/

Any conditions attached:

Distribution of these resources will be free of charge to all participants in the German Human Genome Project, otherwise charges will be made to cover manufacturing expenses and postage costs.

In the case of some libraries additional conditions governing usage and distribution have been imposed by the owners.

RESOURCES AVAILABLE

Name of participant : DAVID J. LIPMAN

Nature of resources available (software, maps, clones etc.)

Databases & Software
see : http://www.ncbi.nlm.nih.gov

Available via: WWW, FTP, CDROM

Any conditions attached: none

RESOURCES AVAILABLE

Name of participant:

Dr. Robert K. Moyzis
Center for Human Genome Studies
Los Alamos National Laboratory
Los Alamos, New Mexico 87545

Ph[redacted]

Nature of resources available (software, maps, clones, etc.)

A) Complete digest libraries for each human chromosome
B) Partial digest phage and cosmid libraries for approximately half of the human karyotype (phage: 4, 5, 6, 8, 11, 13, 16, 17, X; cosmid: 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 20, X, Y)
C) YAC libraries for human chromosomes 9, 12, 16 and 21
D) M13/STS libraries (can be constructed for any human chromosome)
E) High-resolution YAC/STS/cosmid maps of human chromosomes 5 and 16

Available via:

A) American Type Culture Collection
B) Request from Los Alamos. Will also be available from commercial sources
C) Request from Los Alamos
D) Collaboration with Los Alamos
E) htpp://www-ls.lanl.gov; GDB and GSDB; request materials from Los Alamos

Any conditions attached:

A) Small fee; agreement to acknowledge Los Alamos in publications
B) Must sign Material Transfer Agreement with University of California limiting use to scientific purposes, limiting further distribution and agreeing to a limited collaboration with Los Alamos investigators
C) Collaboration with Los Alamos
D) Collaboration with Los Alamos
E) Sequencing coordinated with Los Alamos

RESOURCES AVAILABLE

Name of participant : Richard Myers + David Cox

Nature of resources available (software, maps, clones etc.)

- two panels of whole genome radiation hybrid DNAs
  (Stanford G3 panel - 400 kb resolution)
  (Stanford TNG panel - 100 kb resolution)
  available from Research Genetics
- map positions of 7300 STSs on the G3 radiation hybrids
- an email server allowing anonymous STS radiation hybrid scores to be integrated ~~on the~~ with our mapping data on the G3 hybrids

Available via: http://www-shgc.stanford.edu

Any conditions attached:

- none

RESOURCES AVAILABLE

Name of participant : Bruce Roe

Nature of resources available (software, maps, clones etc.)

Laboratory Protocols

Cosmid, P1 and BAC's sequence data (IN progress)

Available via: HTTP://dna1.chem.uoknor.edu

Any conditions attached:

let us know if you find something cool that we missed

RESOURCES AVAILABLE

Name of participant : Melvin I. Simon

Nature of resources available (software, maps, clones etc.)

1. Mouse 129ES Cell - BAC Library 235,000 clones (~10X coverage)
2. Human Fibroblast - BAC Library B 70,000 clones (~3X coverage)
3. Human Sperm BAC Library C 75,000 clones (~3X coverage)
4. Human Primary Fib. BAC Library A 100,000 clones (~4X coverage)
5. Human Sperm BAC Library D 75,000 clones
6. 619 - Ch 22 Specific Mapped BAC clones

Available via:

1, 2 and 3 Now Available - Research Genetics Inc (Huntsville Ala)
5 & 6 Available - Research Genetics Inc (April 1996)
4 - Available for screening via Hiroaki Shizuya -
Biology Division Caltech - PASADENA
FAX - (818) 796-7066
Also see:
http://www.tree.caltech.edu

Any conditions attached:

No conditions or restrictions are attached to this material.

RESOURCES AVAILABLE

Name of participant :

Jim Weber

Nature of resources available (software, maps, clones etc.)

Crude, but comprehensive human linkage maps

STRP information

Methods

Image analysis software

Construction information for water bath thermal cycler and some SCAFUD components

Sequence assembly simulation program (from Gene Myers at University of Arizona)

Available via:

Website: http://genetics.mfldclin.edu
Email: gene@cs.arizona.edu

Any conditions attached:

Software is not supported.

RESOURCES AVAILABLE

Name of participant :

Jean Weissenbach

Nature of resources available (software, maps, clones etc.)

The Généthon Human Linkage Map
(5,264 microsatellite markers)

Map + description of reagents
(sequences, primers, alleles, frequencies, etc.)

Available via:

http://www.genethon.fr

Any conditions attached:

freely available

Bacterial genome projects at TIGR

Organism	Status	Size (Mbp)
Haemophilus influenzae	Science 269:496, 1995	1.80
Mycoplasma genitalium	Science 270:397, 1995	0.58
Methanococcus jannashii	completed in '95, publication spring '96	1.7
Streptococcus pneumoniae	in progress, to be completed by early '97	2.3
Helicobacter pylori	in progress, to be completed by late '96	1.8
Treponema pallidum	funded by NIAID, to be completed by late '97	1.0
Archaeoglobus fulgidus	funded by DOE, to be completed by early '97	2.0
Deinococcus radiodurans	funded by DOE, to be completed in '98	3.0
Mycobacterium tuberculosis	pending at NIAID	4.2
Vibrio cholera	pending at NIAID	2.0
Porphyromonas gingivalis	pending at NIDR	2.2
Neisseria meningitidis	discussions with Wellcome Trust	1.9
Borellia burgdorfori	pending at private foundation	1.3

Other genomes at TIGR

Arabidopsis thaliana	pending at NSF — 3 years, 7 Mbp	20-30
Human	pending at NCHGR — 3 years, 30 Mbp	30
Plasmodium falciparum	pending at DOD	30
Plasmodium vivax	pending at DOD	30

THE WELLCOME TRUST

183 Euston Road
London NW1 2BE

[redacted]

Fax: [redacted]

E.mail [redacted]

JS/SO'D/LET560

12th February 1996

Dr Dr Peterson
National Institutes of Health
National Centre for Human Genome Research
38 Librry Drive MSC 605
Building 38a Room 610
Bethesda MD 20892-6050
USA

Dear Dr Peterson

International Strategy Meeting on Human Genome Sequencing
25th-28th February 1996

I am writing to you with final instructions and arrangements in respect of your attendance at the meeting on Human Genome Sequencing to be held at The Hamilton Princess Hotel, Bermuda from the 25th to 28th February 1996. Please find enclosed a copy of the final programme. Included at the end of the programme is a page entitled "Resources Available". **This should be completed prior to the meeting and be handed to me on arrival at the hotel** and will be discussed in Session 2.

A brochure of the hotel is attached, and contact details at the hotel are as follows:

The Hamilton Princess Hotel
P.O.Box HM 837
Hamilton HM CX
Bermuda

Tel: 809 295 3000
Fax: 809 295 1914

Ground transportation to and from the airport has been arranged with Bee-Line Transportation who have been sent details of all flight arrivals and departurnes. Please let me know **immediately** if your flight schedules have changed from those stated on your registration form.

Accommodation has been arranged for you at The Hamilton Princess Hotel for the nights of 25th, 26th and 27th. During the meeting the Trust will meet the cost of your accommodation and subsistence, plus the drinks receptions and wine at the conference dinner on Monday evening as detailed in the programme. All other drinks, telephone calls, newspapers and other incidentals should be met by individuals prior to departure from the hotel. The Trust does not therefore expect to meet the cost of any additional expenses. Could we also ask that, for ease of handling at the hotel, **all your luggage is clearly marked with your name**.

Please note that it is a requirements of The Hamilton Princess that for all evening functions dress for delegates should be smart casual with gentlemen wearing a jacket. This dress code also applies to the restaurant for the conference dinner. The dinners, including the conference dinner, are an essential part of the programme and delegates are therefore expected to attend all of these events unless previously agreed with the Trust.

It is the policy of the Trust that all delegates are expected to stay for the entirety of the meeting unless personally agreed with the Trust prior to the start of the meeting. Unfortunately, due to recent experiences, it is necessary to add that no unexpected delegates will be accepted to take part in the meeting.

In the event of severe delays on your way to the meeting or any last minute changes to itinerary please contact me as soon as possible. I may be contacted at The Hamilton Princess from Thursday evening, 22nd February 1996 on telephone number 441-295-3000 or facsimile 441-295-1914.

I look forward to meeting you at The Hamilton Princess later in the month, and to an interesting and sucessful conference. In the meantime, should you require any further assistance, please do not hesitate to contact me.

Yours sincerely

Jilly Steward
Meetings and Travel Manager

Third International Strategy Meeting on Human Genome Sequencing

Hamilton Princess Hotel, Bermuda,
27th February - 1st March 1998

Sponsored by

The Wellcome Trust

US National Institutes of Health

US Department of Energy

164th Genome Technology Committee,
Japan Society for Promotion of Science

UK Medical Research Council

SUMMARY

PROGRESS

- A total of 116 Mb of finished human genomic sequence was reported, with over 107 Mb submitted to GenBank/EMBL/DDBJ (see table 1)
- This represents approximately 3% of the human genome
- More than 70 Mb had been produced in the preceding year
- The Sanger Centre and Washington University had finished respectively 21.4 Mb and 19.7 Mb of human genomic sequence in the preceding year. Six other centres had finished 3-7 Mb
- Most centres reported that production was being ramped up steeply although last years' sequence production *in toto* had not met the predicted output by a factor of 2.1
- Generating a sequence-ready map was no longer a limiting factor in some centres but globally matching clone production to sequencing capacity was still an issue
- Finishing of sequence i.e. closing gaps in the map and reading through difficult sequence was identified as the major bottleneck
- Other factors limiting scale-up of the sequencing effort were staff, space and funding.

QUALITY AND COSTS

- The previously agreed quality standards were endorsed
- During the ramp-up phase cost reductions were unlikely if the quality standards were to be maintained
- All attendees agreed to participate in an International Quality Assessment exercise, this would be based on the exchange and reassembly of raw data.
- Closing gaps in the sequence will be a significant cost and therefore contiguous sequence should be generated as soon as possible in order that the full costs of achieving contiguity are confronted and not deferred
- Groups were asked to start reporting contiguity of data
- Eventually groups should only report sequence in contigs over 500 Kb
- Working groups were to be convened to explore the following issues:
    - consortium buying
    - cost accounting
    - clone fidelity
    - finishing practices.

DATA RELEASE

- US, UK and German funding agencies confirmed that their policies were consistent with the principle of immediate data release for high throughput human genome sequence.
- Most US, UK and German centres reported achieving or aspiring to the target of immediate data release
- The Japanese Science and Technology Corporation does not restrict or require immediate data release, however finished data has to be released via the JST
- GENOSCOPE (France) - data from 'in house' projects would be released immediately, no such requirement was being made for collaborative projects
- Participants offered their support to ensure early data release for all human high throughput sequencing projects
- Collaborations should be encouraged with groups with biological or sequencing interests within targets, but the principle of early data should apply to all sequence
- Participants agreed to apply the same conditions of early data release to mouse genomic sequence
- Participants urged high throughput sequencing projects on all organisms to adopt a policy of immediate data release
- The policy of early data release was agreed to facilitate the co-ordination of the project and to allow researchers early access to the sequence. To ensure that the unfinished sequence was accessible to the community sequencers were encouraged to
    - submit their unfinished sequencing data to the unfinished divisions of GenBank, EMBL, DDBJ
    - develop chromosome specific Web sites
    - ensure presentation of data on individual Web sites was user friendly.

HUMAN SEQUENCE MAP INDEX

- All participants confirmed their support for a single World Wide Web site containing sequencing interests
- Regions defined on the Index should be viewed as expressions of interest rather than exclusive claims
- All groups undertook to submit their interests to the new Index at the National Center for Biotechnology Information
- Several participants believed that the limits agreed at the last meeting on the duration and size of stated interests were too restrictive in the context of the current ramp-up of sequence production
- Mapping and sequencing targets needed to be shown separately on the Index
- Sequencing targets should include short term goals
- In the context of claims registered on the Index being understood as expressions of interest, mapping targets could include regions of longer term interest for which map data were being generated
- The Genethon markers available on the Index were not sufficient to accurately define regions of interest. David Cox (Stanford University) and David Bentley (Sanger Centre) were assigned authority to advise the Index on additional markers and to place these on the consensus map if necessary
- A simple visual representation of sequencing centres' target areas would be added to the Index.

SEQUENCE-READY MAPS

- A small number of BAC libraries had been made and distributed from DNA obtained in line with ethical guidelines for large scale sequencing; several more libraries would be produced
- YAC libraries made from a similar source of DNA may be required
- Some sequencers expressed an interest in a large insert BAC library for gap closure
- The possibility of a mapping consortium for the mouse received support.

ANNOTATION

- The participants were agreed on the importance of annotation to exploit the value of genomic sequence
- A survey of annotation in the sequence databases demonstrated that sequencing centres were annotating sequence to different levels and in different manners
- There were differing opinions as to the level which sequencing centres should annotate sequence themselves
- It was agreed that a minimum level of annotation for sequence submissions to the sequence databases would be required
- A method for displaying accuracy data which would be of use to sequence users would be developed
- A more standardised system of annotation should be considered with features described in a searchable and consistent manner
- Including the evidence for the annotated feature should be considered
- A system of cataloguing the sequence submissions would be developed. Summary information for each contig would contain information on the contig and constituent clones.

NEW TECHNOLOGY

- All centres were still reliant on existing technologies although several were investigating improvements in sequencers, enzyme and dye technology
- Interaction between centres to share innovations should be encouraged
- It was unfeasible to try to change technology at same time as ramping-up sequence production.

NEXT MEETING

- The location to alternate between UK and US, at a site convenient for major airports.
- The working groups to report their findings
- A similar number of participants to be invited.

INTRODUCTION

Michael Morgan welcomed participants to the Third International Strategy Meeting on Human Genome Sequencing and thanked the other sponsors of the meeting for their support.

Session I PROGRESS, STRATEGIES AND DEVELOPMENTS
Chair: James Watson

James Watson opened the first session by commenting on the budget and timeline for the human genome sequencing project. It was becoming apparent that completion of the project by 2005 would probably cost more than had been originally envisaged. For successful completion of the project in 2005, it was critical to maintain momentum in order to retain both staff and resources. There was a need to promote a long term vision of genome sequencing as an activity integral to scientific discovery. The genome sequencing community should aim to convince both the scientific community and the public of the challenges involved in the project, and value of genome sequence information. Only in this way would support be obtainable to build a superstructure to maintain genome sequencing as a long term activity and to sequence the mouse, *Drosophila, Fugu* and other complex genomes.

PROGRESS

All the sequencing centres had been asked to provide information by email on progress made to date and this, together with additional information on strategies and predicted sequence production provided by the groups at the meeting, is contained in Appendix 1.

It was reported that, to date, 600 Mb of human sequence has been deposited in the sequence databases GenBank, EMBL and DDBJ; the major contributor being EST sequences. The total finished human genomic sequence reported was 116 Mb with over 107 Mb submitted to the databases. This represented approximately 3% of the human genome. Two centres had finished more than 7 Mb of human genomic sequence in the preceding year: the Sanger Centre and Washington University. The human genome sequencing capacity of both these centres would increase with the completion of the *C.elegans* genome in 1998. Six other centres had produced between 3 and 7 Mb during the past year (Table 1).

Most of the sequencing centres represented at the meeting reported that they were in the process of ramping-up their sequence production very steeply. Most groups had over predicted sequence output for the preceding year compared to what had actually been achieved, although the 70 Mb plus produced in the past year exceeded the cumulative total reported a year ago (Table 1). It was questioned if the ramp-ups predicted by the groups were actually feasible in the context of the previous year's production; most producers were confident that their goals were feasible, if ambitious. The groups cited staff recruitment and retention, space and funding as critical factors in achieving the increase in sequence production.

The recruitment, training and retention of high quality staff at all levels, particularly in the finishing process, was proving to be one of the greatest challenges. The project needed to recruit high quality staff as technology innovators, process managers and technicians. Currently this was proving to be problematical for several reasons. The higher salaries offered by biotechnology companies were more attractive, particularly for the technical staff.

Individuals of the calibre sought as innovators also had the option of pursing a more biologically oriented academic career. It was accepted that under the present system a career in genome sequencing was unlikely to result in academic progression. The group at the University of Washington reported that the only solution they had found to provide appropriate career opportunities was by moving away from the university system and forming a company for genomic sequencing. Some individuals could possibly be attracted by combining sequencing with the opportunities for biological research, however, many groups had found sequencing too demanding to effectively combine with other activities. A more appropriate career structure was required which would recognise the unique skills required and provide appropriate rewards. If the long term future of genome sequencing was more secure, as had been discussed above, it was thought that there would be more individuals prepared to commit themselves.

As well as improving career prospects the profile of genome sequencing needed to be raised within the university community, and this had been attempted through seminars. It was reported that, at least at Baylor, there had recently been an increased interest in pursuing a career in the area.

STRATEGIES

Mapping and Sequencing

Most groups involved in producing their own sequence-ready maps reported now being able to produce sequence-ready clones at a sufficient rate to satisfy their current sequencing capacity. Those groups not involved in mapping, or without sufficient mapping capacity, were hoping that other genome centres, as well as other members of the scientific community, would continue to provide them with sequence-ready maps. There were concerns expressed about the quality of some of the maps provided, and most centres without any mapping capacity identified a need to develop some mapping capacity in order to validate sequence-ready maps both prior and post sequencing. Therefore globally there were still concerns as to whether the mapping capability would be sufficient to satisfy the rapidly increasing sequencing capacity.

There was some variation between centres in their strategies and priorities. Most groups were now using or moving towards BAC libraries and shotgun sequencing. Strategies for generating sequence-ready maps were discussed in more detail in session IV but there were differences in emphasis with groups concentrating on achieving contiguity or coverage early on. Groups aiming for early contiguity were either ordering STSs prior to generating clones or using a strategy including directed generation of additional probes from YACs to close gaps. Most groups reported an average contig length for finished sequence in the range 130-220 Kb, with largest contigs in the 1 Mb range.

Finishing

Most centres identified finishing sequence as the major bottleneck in the production process. This was due to both a lack of personnel and the difficulty in closing gaps owing to both cloning and sequencing difficulties. There were moves towards increasing the automation of the finishing process to enable finishers to concentrate on truly problem areas. A semi-automated prefinishing step using primer walking was being used at UTSW and Oklahoma University. The process was reported to be cost effective owing to the MERMAID oligonucleotide sequencer, which was built in house, and could produce oligonucleotides extremely cheaply. More systematic methods of closing difficult gaps both in the map and sequence were also required. New technologies and a mechanism for sharing finishing practices were discussed in more detail in later sessions.

Table 1

FINISHED HUMAN GENOME SEQUENCE (Mb)

Centre	Cumulative Output Feb 1997	Cumulativ e Output Feb 1998	Total in Sequence Databases Feb 1998	Actual Output Feb 1997-8	Output Feb 1997-8 -as predicted Feb 1997
Sanger Centre	14.6	36.02	34.85	21.4	35
Washington University	4.8	24.5	22.6	19.7	24
Whitehead Institute	0.08*	7.0	6.6	6.9	20
TIGR	2.7	6.4	6.4	3.7	11
Baylor College Of Medicine	3.0	6.5	5.9	3.5	12
University of Washington	0.59	3.65	3.65	3.06	6
ACGT - ABI	2.4	5.0	5.0	2.6	3.5
DOE Joint Genome Initiative	0.96**	7.67**	5.67**	4.71	20
IMB - Jena	1.5	5.29	5.29	3.79	6
UT Southwestern	1.6	4.35	4.35	2.75	5
ACGT University of Oklahoma	3.8	4.13	4.13	0.3	5.5
University of Tokyo	2.7	5.1	2.9	2.4	3.4
Stanford University	0.3	0.65	0.51	0.35	5
TOTALS	**39.03**	**116.26**	**107.85**	**75.16**	**156.4**

Figures for Cumulative Output February 1997 and Predicted Output February 1997-8, taken from the Report of the Second International Strategy Meeting (unless indicated otherwise).

Figures for Cumulative Output February 1998 and Total in Sequence Databases February 1998 extracted from individual reports in Appendix 1.

* Total revised down from 2.1 Mb reported at the Second International Strategy Meeting. 2.1 Mb did not correspond to finished deposited sequence but rather sequence in progress

** Revised totals submitted after the Third International Strategy Meeting. Total for February 1997 of 4 Mb revised down to 0.96 Mb to include only sequence finished to agreed quality standard.

Session II SEQUENCING QUALITY AND COSTS
Chair: Eric Lander

The aim of the session was not to discuss actual figures for quality and costs but to put in hand a process that would be able to determine these with a greater degree of confidence than had been possible previously.

The Human Genome Sequence produced should be characterised by the "four As"

- Accurate (10^{-4})
- Assembled (> 500 Kb contigs - no gaps)
- Affordable
- Accessible

Quality and cost had been discussed in December at the NHGRI Principal Investigators' meeting. At that meeting it was apparent that it would not be possible to reduce costs during the ramp-up phase of the project, as had previously been envisaged, if the agreed sequence quality standards were to be met.

SEQUENCE QUALITY

Accuracy

The standards agreed in previous years for sequence accuracy were still felt to be appropriate. It was also recognised that there was a need for a mechanism to demonstrate that the sequence produced met these standards. In the past year there had been two large scale sequencing Quality Assessment exercises involving NHGRI funded centres. The mechanism suggested at the last strategy meeting had been used, involving exchange and reassembly of raw data. There was unanimous support for this exercise, both as a means of sequence quality control and for identifying how errors occur. The majority of clones reviewed in the last exercise had met the required quality standards (Appendix 2a). Those groups with sequence below the required accuracy standard of 1 error in 10,000 bp had re-analysed their sequence to ensure it met this standard.

The results of the internal QA exercise at the Sanger Centre were reported (Appendix 2b). The quality of the inspected sequence had been well above the required standard. Errors detected were mostly associated were older data. It was reported that a strategy of allowing a controlled level of uncertainty in problem sequence had been introduced as alternative to gaps in the sequence. After all current methods of reading through a region were exhausted, a 'black-tag' could be applied to identify regions where the confidence in the sequence was lower than required. The tag would remain on the annotated sequence until new methods were developed to successfully read through the region. The frequency of the tags was highly variable, but a rough estimate was given as 1 every 100 Kb.

All participants were enthusiastic about participating in an International QA exercise modelled on the two NHGRI exercises. As before, raw data would be exchanged between sites for reassembly. The focus would be on clones sequenced since September 1997, as sequence produced before this date had been sampled in previous exercises. There was some discussion on the precise format for the exercise, in particular the need to exchange reagents as well as the raw data, the level of checking required by each centre, and the inclusion of overlapping clones to check the fidelity of clones to the target region. Participants in the last QA exercise had found the access to reagents very useful when ambiguities/errors were detected. The level of checking required by each centre would reflect the relative sizes of the centres. It was agreed that all participants would be on an equal footing and there would be an opportunity for all centres to participate in setting the format and rules for the exercise. The major exercise would begin after 9th October, the date of submission of NHGRI grant applications, but would be completed in time for the review of the applications. In addition there may be a partial exercise before 9th October: the results of which would be available before 9th October.

Following the December meeting, the ability to submit accuracy data in the form of a cumulative PHRAP curve was being added to GenBank. There was a suggestion that providing the following information would give a truer indication of the quality of the sequence.

1. Cumulative error
2. Distribution of poor quality bases
3. Consensus changing edits
4. List of poor quality bases

The QA assessment had highlighted the value of co-operation between centres for sharing information on ways of improving the process and solving problems. Centres were using differing strategies for assembling and finishing data, although mostly based on PHRED, PHRAP, and XGAP and/or CONSED. A working group would be set up to cross compare procedures and possibly identify the most effective practices. The group would also examine the issue of difficult gaps, how these could be identified prospectively, and the methods groups had used to close them.

A number of errors detected in the second QA exercise had involved clone instabilities. The reasons for deletions in clones and mechanisms for identifying and avoiding them were discussed. There was no standard measure for clone validity, in particular how many consistent clones had to be isolated to give an acceptable level of confidence. Some groups were worried about the possibility of systematic deletions in M13 and several groups stated they were now concentrating on pUCs for subcloning. However, most deletions found in multiple subclones were thought to arise from deletions in the source BAC/cosmid. Deletions in the source clone could generally be detected using the fingerprint analysis to spot subpopulations of deleted clones. It was agreed that clone fidelity would be considered further by a second working group.

Contiguity

The issue of contiguity of sequence had been discussed extensively at the December meeting of the NHGRI investigators. It was accepted that the sequence produced by the human genome project should be contiguous, however, there were some differences of opinion as to how and when this should be achieved. The compelling reason for achieving contiguity early on was the potentially huge deferred cost involved if gaps had to be closed at a later date. Some groups were using mapping and sequencing strategies aiming to generate contiguity in parallel with the sequence production, but other operations were initially concentrating on good coverage. Larger operations were using chromosome wide mapping strategies, with sequencing efforts being nucleated from several sites as the map was extended. These centres were concerned that the drive for contiguity should not prevent the most optimal mapping strategy being used.

There was a move to encourage contiguity of submitted sequence by imposing a limit on reported 'finished' sequence to that in contigs greater than 500 Kb. It was agreed that this was not be feasible at the present moment and effort should not be overly diverted from achieving the necessary sequence throughput. In order to encourage sequence contiguity there should be a mechanism to judge contiguity and to give credit to those groups with a greater emphasis on contiguity as they would be generating less unique sequence. It was agreed that the number average (amount of sequence submitted/number of sequences submitted) gave a better idea of the number of gaps in sequence than a system weighted by the distribution of contig lengths.

COSTS

It was generally accepted that the cost/bp was unlikely to drop in the fashion predicted at the previous meeting. If quality standards were to be maintained the cost/bp would probably stay static at about 50 cents/base during the next three years as sequence production was ramped-up. It was reported that the NHGRI budget for the current fiscal year for production sequencing was $70 million, with requests from current centres totalling $85 million. Following recent budgeting trends, it was possible that as much as $140 million per year may be available within 4 years. Therefore if costs started to fall after three years at the rate of 10%/year the NHGRI would still be able to meet its targets but not with a great deal of flexibility.

There was no support from the sequencers for exploring whether reducing the accuracy of sequence below 10^{-4} could significantly reduce costs. There was agreement that it would be misleading to assume a simple correlation between cost/bp and accuracy, as the high degree of accuracy had additional implications such as facilitating sequence assembly.

The reasons for cost assessment were as follows:

1. Identify who is cheapest - although this would be inappropriate in the ramp-up phase
2. Process improvements - to target savings at the most effective point
3. Cost comparison - to share cost effective methods and best practice
4. Cost projection.

In the first session there had been an attempt to estimate the amount of space and number of people involved in each centre per Mb of sequence produced. This had illustrated the difficulty involved in trying to cross compare operations. It would be impossible to compare costs between centres exactly but it could be possible, as was possible in business, to sub-classify costs sufficiently to make comparisons meaningful. A working group was agreed to consider if it was possible to devise a system for making cost comparisons that would be useful to groups. The earliest it could be expected that a workable system would be in place was next year.

The major cost drivers were considered to be salaries and sequencing reagents. In view of the genome community being a significant purchaser of sequencing reagents the community should be able to negotiate with suppliers from a position of strength. Funding agencies, such as the NIH, could not be involved in consortium buying but a group would be formed to investigate this further.

SUMMARY

An international QA exercise would be undertaken.
Working groups would be formed to report at the next strategy meeting. By the conclusion of the meeting the composition of working groups had been agreed as follows:

Finishing
Bob Waterston (chair)
Phil Green
Sanger Centre representative
Eric Lander
Ellson Chen

Clone Fidelity
David Cox (chair)
Phil Green
John MacPherson
Bruce Birren

Cost Accounting
Eric Green (chair)
Jane Rogers

Consortium Buying
Rick Wilson (chair)
Rick Myers

The optimal size would be 5-6 people per group and volunteers were invited. Sources of expertise outside of the academic community should be utilised if appropriate.

Session III DATA RELEASE AND AVAILABILITY
Chair: Richard Gibbs

FINISHED SEQUENCE IN THE SEQUENCE DATABASES

The amount of sequence released into finished division of GenBank by the genome sequencing centres had been calculated. There was currently no single method to do this and therefore three methods had been used.

1. Sequence submitted using the direct ftp route for high throughput genome sequence - total 100 Mb.
2. Finished sequence labelled as htg (high throughput genome) sequence - total 60 Mb.
3. Human genome sequence records >10 Kb (manually assessed to fit certain criteria) - total 100 Mb.

For various reasons none of these methods were completely satisfactory for calculating the amount of finished sequence submitted to the sequence databases. Four suggestions were made to facilitate this in future.

1. Centres to maintain a simple list of accession numbers for sequence submitted to the finished division of the databases

2. A similar list to 1 but clones grouped under the appropriate contig name

3. A similar list to 2 but with information on how the sequence could be assembled from individual clones

4. A description of each contig including size, orientation, position and source of each clone and how each clone fits into the contig.

There was a consensus that a list of sequences submitted to the databases would be valuable. An option providing information on the contig assembly was supported as the current situation, with the assembly information was embedded in sequence records, was less than satisfactory. It was noted that a similar system was working effectively for the *C.elegans* sequencing project. GenBank/EMBL volunteered to put together a Web site containing this information and also to download the information if it was provided on each centre's Web site. To ensure that the information provided was meaningful, centres should update the information regularly. Jim Ostell undertook to circulate to the groups a more detailed proposal based on option 3 for comments in the near future.

ACCESSIBILITY OF SEQUENCE DATA

The rationale for releasing unfinished sequence data in to the public domain was to make it accessible to the scientific community at the earliest opportunity as well for co-ordination of the project. Most unfinished data were available from centres' Web sites although the sequence databases now provide a division for unfinished sequence. To ensure that the unfinished sequence was accessible to the community, sequencers were encouraged to submit their unfinished sequencing data to the databases' unfinished divisions. If data were only being released on Web, the possibility of developing a joint Web page for unfinished sequence on each chromosome, similar to those for chromosomes 21 and 22 should be considered. Centres should also ensure that sequence information available through their own Web sites was in an accessible form. André Rosenthal's group in Jena had visited and assessed each site from the point of view of the user community. Two recommendations to increase the accessibility of the data were made:

1. sequence and the corresponding maps should always be cross referenced
2. the capacity to perform BLAST searches on the sequence data was absolutely necessary.

DATA RELEASE POLICY

The first Bermuda meeting had agreed to a policy of immediate release of human genome sequence produced by high throughput centres. The funding agencies present were asked to state whether this policy was being adhered to. The policy of the US agencies (NIH and DOE) was to require its grantees to comply with the Bermuda principles. Some centres had not achieved full compliance as yet, but were taking steps to do so. Posting assemblies greater than 2 Kb on the Web within 24 hours was satisfactory, but submission to the unfinished division of GenBank was preferred.

The Wellcome Trust reaffirmed its commitment, and that of the Sanger Centre, to early data release. Confirmation from the BMBF that the German genome sequencing consortium could now fully adhere to the Bermuda principles was warmly welcomed.

The French genome sequencing effort was in its initial stages: 'in-house' projects would target human chromosomes 3 and 14 and data would be released immediately. Tenders for collaborative projects would be considered 2-3 times a year by the French ministry. Data release for collaborative projects would be considered on a case by case basis. However the Bermuda principles applied to all sequence produced by high throughput centres and, although there was no human collaborative projects at present, the participants were worried by future problems in adhering to the principles. The possibility of refusing projects where collaborators requested restrictions on data release had not been explored as yet. The participants offered their assistance to change this policy as support from the sequencing community had helped remove the restrictions on data release originally required by the BMBF.

It was reported that the Japanese Science and Technology Corporation (JST) do not require or restrict their researchers from releasing unfinished data immediately. It was noted that Yoshiyuki Sakaki was releasing his unfinished data on to the Web and to DDBJ almost daily. The problem arose with the requirement to release finished data via the JST database which resulted in a delay of approximately three months in the release of data. The participants were unable to appreciate the rationale for requiring finished data to be released via the JST and the consequent delay in the release of finished sequence.

The discussion on data release was resumed prior to the final session, in the context of a policy for genomic sequence from other organisms.

Session IV ALLOCATION OF REGIONS / ETIQUETTE FOR SHARING
Chair: David Cox

The session revisited the conclusions on sequencing claims and etiquette from the Second Strategy meeting. There was agreement on the following points from the summary of the previous year's meeting:

- mapping investment did not automatically entitle sequence claims over the region
- potential conflicts between sequencers should be resolved by early communication
- collaborations with groups with a biological interest in the region sequenced should be subject to the same principles of data release and communication.

HUMAN SEQUENCE MAP INDEX

Relocation of the Index

Following the discussion at last year's meeting, the Human Sequence Map Index had been relocated from HUGO to NCBI. There was unanimous agreement that the Index was the appropriate mechanism to publicize sequencing activities and that all high throughput sequencers would submit their sequencing interests to the Web Index. Those centres which had not submitted information to the Index gave a lack of suitable markers to define their regions of interest as the limiting factor.

Two proposals to improve the Web site received general support.

1. The Genethon markers available on the Index were not sufficient to accurately define regions of interest. An expanded set of 2,000 markers and telomeres would be added to the Index; possibly as many as 3,000 should be available to give a density of 1 marker per 100 Kb. David Cox (Stanford University) and David Bentley (Sanger Centre) were assigned authority to advise the Index on additional markers and to place these on the consensus map if necessary.

2. A visual representation of sequencing activities using ideograms of chromosomes had been produced by the Sanger Centre and Washington University. A similar idea would be incorporated in the Index and possibly on other sequencing centres' home Web sites. This would provide a very useful summary of regions being targeted. Regions of sequencing and mapping interest as well as regions of the finished sequence could be represented on the Web Index in this manner.

Submissions to the Index

The Index should not be seen as a mechanism to enable exclusive claims to be made, but rather to be a source of information of regions being targeted by the sequencers to help facilitate co-ordination and collaboration between sequencing centres and the rest of the scientific community. Generally the group in the best position to sequence a region should be given priority.

Centres appeared to be using different criteria to define regions of interest on the Index. It was proposed that the limit on the size of the regions indicated as three times the sequence output of the previous year was too restrictive in the context of the current ramp-up of production. Also the time limit of one year prior to sequence production did not take into account that mapping needed to start well in advance of sequence production to ensure clone supply. There was extensive discussion as to what it was reasonable to claim as a target. There was a difference between centres concentrating on mapping smaller very defined regions with an emphasis on obtaining contiguity as soon as possible and the centres involved in chromosome wide mapping strategies. There was concern that allowing larger claims and longer term interests on the Index would reduce the pressure to produce contiguous sequence. However the larger centres were concerned that not allowing a chromosome wide (ptel-qtel) 'claim' on the Index would not reflect their chromosome wide mapping strategies and could either lead to a less optimal mapping strategy having to be used or conflicts arising because of a lack of information on the Index.

It was felt that by separating sequencing activities from mapping interests the Index could provide the most accurate reflection of the sequencing groups' activities and address the concerns of centres about the extent of claims. Sequencing interests should define the short term goals of centres: this was essential information as it was at this point that there was very little flexibility and a high level of commitment had been made. It was noted with interest that the JGI were providing on their Web site a schedule for the regions being sequenced. This provided a very transparent view of the sequencing interests at the JGI and demonstrated the level of commitment to a region.

Mapping interests should be indicated separately from sequencing interests on the Index, and should define those regions of longer term interest for which mapping was being undertaken. The extent of regions defined would depend on the mapping strategy used.

FULFILLING COMMITMENTS

There was a need to ensure that sequencers were fulfilling their commitments by producing contiguous sequences for the regions 'claimed' on the Index. This was especially important if the previous limits were no longer be imposed on duration and size of 'claims' and where groups were undertaking chromosome wide mapping strategies. The Sanger Centre explained that in stating a chromosome wide interest in chromosome 1 it was signing up to take responsibility for obtaining the contiguous sequence from the chromosome. If smaller groups expressed interests in particular regions the Centre would be more than willing to take these interests into account in their sequencing strategy. This approach was felt to be helpful towards smaller groups. The strategy meetings would serve as a forum for sequencers to review whether groups were fulfilling their commitments to the regions claimed by producing contiguous sequence.

Session V SEQUENCE-READY MAPS AND RESOURCES
Chair: Eric Green

From previous discussions during the meeting it was obvious that the type of mapping strategy undertaken had a profound influence on many aspects of a sequencing project. The production of sequence-ready maps could be seen as a two-step process analogous to the sequencing process, with a shotgun stage (the initial screens for clones) and a finishing stage (obtaining clone contiguity). Similar to sequencing, it was the finishing process (i.e., obtaining complete coverage of the region) that was proving to be the more problematical task.

RESOURCES

It was reported that two BAC libraries had been produced by Mel Simon at Caltech. There was limited capacity at Caltech for further library validation and filter production. The libraries are available from Caltech and via Research Genetics.

Pieter de Jong confirmed that male and female human DNA had been obtained in an IRB-approved manner for use in large-scale sequencing. The male BAC library (RPCI11) had been made and distributed world-wide. The library is divided into four segments, with an overall genome coverage of 25X, an average insert size of 174 Kb, 1.2% empty wells, and 0.8% non-recombinant clones. A second library from the same donor with larger inserts (230-240 Kb) has also been made; coverage was only 1X but could be expanded if it proved useful. A human female library is also in the pipeline and will be available by early summer. Three mouse libraries have been made and were being distributed, the best being the C57BL/6J library (RPCI23). All libraries (including high-density filters) are available through RPCI, Research Genetics, the MRC HGMP Resource Centre, and the Max-Planck Institute. Libraries from a variety of other species are also being constructed; all would be in the public domain irrespective of the funder.

The libraries are being validated using the following criteria agreed at the NHGRI PI meeting:

- Clones analysed for average insert size
- % empty wells
- Probing the library with >10 unique probes
- Probing with STSs at 5-10 Kb intervals for 2-3 1 Mb regions and fingerprinting the resulting clones.

One 'anomaly' that was reported between libraries was that for certain probes there is a representation bias dependent on the vector system and restriction enzyme.

TIGR and Leroy Hood are in the process of deriving 600,000 BAC end sequences. 60,000 have so far been sequenced. The completion of the project is scheduled for the end of September 1999. The end sequences data are being updated on the TIGR Web site nightly and were being submitted to GenBank on a weekly basis (appendix 1 - TIGR submission).

Most of the centres reported that they are using Pieter de Jong's libraries as the major source of clones for sequencing and they are very happy with the quality of the library and filters. The technique of choice for identifying clones is hybridisation to filters. There seemed to be only limited need for PCR pools. PCR pools for one or two libraries could be useful, as PCR data could be used for resolving ambiguous hybridisation data. For pools to be truly useful, they would need to be of a higher quality than the commercial pools currently available.

Many centres are very interested in the possibility of one good deep, large insert BAC library, particularly for closure of mapping gaps. It was thought that with current technology, the maximum insert size for BACs is around 300 Kb. The limiting factor is electroporation, but this could possibly be circumvented by using a T4 packaging system. This would be explored further if there was a high level of interest from the sequencing community. Pieter de Jong is considering producing additional mouse and human libraries using a modified BAC containing a yeast centromere and a HIS3 marker This vector would enable isolation of clones from a different haplotype or from closely related species by homologous recombination in yeast.

It was reported that about 20% of the *C.elegans* sequence would be derived from YAC libraries and potentially a significant amount of the human sequence may also need to be derived from YACs. If this turns out to be the case, then new YAC libraries made with DNA obtained in an IRB-approved manner for large-scale sequencing will be required. The sequencers were in favour of using traditional YAC vectors rather than newer vectors (such as circular YACs) since the stability of the latter has not been confirmed. A small-insert YAC library would be more useful for sequencing but a larger-insert library would facilitate the generation of markers.

OPTIMAL MAPPING STRATEGIES

At Stanford, a high-resolution map is being generated by radiation hybrid mapping, with STSs then selected at an even distribution over a particular region (10 markers/Mb). The aim is to generate even coverage over a region by using evenly spaced markers. The density of markers needed to optimise clone contiguity is being investigated.

At the Sanger Centre, 10 markers/Mb is felt to be insufficient. 15 markers/Mb are being used, although it had not been decided if this was actually the optimal density. There is less emphasis on ordering and ascertaining distribution of the STSs before isolating clones. 60-70% clone coverage has been obtained with 10 marker/Mb and 80-85% with 15 markers/Mb. The aim is to obtain a degree of closure by primary coverage without directed effort before walking to close the gaps. For many groups, screening libraries with STSs is not a limiting factor and therefore all the available markers in the target region are being used (rather than selecting markers based on their distribution on a high-resolution map). The problem for most groups is generating markers in regions not covered by the initial screenings.

BAC end sequencing was felt to be a useful additional mapping resource. It was emphasised that to produce high-confidence maps, single clones could not be relied upon and confirmation of the map was required from redundant coverage and fingerprinting. It was noted that the *Arabidopsis* sequencing project has demonstrated that BAC end sequencing could be an effective mapping technique if the data were well validated.

EXPERIENCE AND COSTS FOR A TYPICAL CHROMOSOME

As yet there is little information and no obvious mechanism for capturing the data on mapping costs. So far, from the experience with chromosome 22, the cost of closing map gaps was higher than the cost of map production. More data on the actual costs of finishing a chromosome would be available next year for chromosome 22. Information will be needed to assess the effectiveness of mapping strategies for more typical chromosomes, where a significant amount of extra mapping will be required.

METRIC FOR MONITORING PROGRESS AND QUALITY

Sequence-ready map validation could divided into:

1. Quality standards for the map and definition of a sequence-ready clone
2. Fidelity to the genome.

These issues would fall within the remit of the working group on clone fidelity. The issue of contiguity is also important in the context of map quality, as contiguity minimised the number of end sequences and therefore the uncertainly associated with non-redundant ends.

POTENTIALLY UNCLONABLE REGIONS

The efforts directed towards trying to obtain maps and sequence through problem regions (such as centromeres and telomeres) was raised. Several groups are involved in investigating the potential of obtaining sequence in these regions. The Sanger Centre is investigating the feasibility of obtaining sequence in the centromeric region of chromosome 10 through a collaborative study. Mapping by indirect methods (such as pulse-field gel analysis) is being used; the aim being to sequence into the different classes and intermittent class of repeats. Telomeres have also been causing some difficulties, in particular deleted clones and short sequencing reads were obtained from DNA near to telomeres. Unpublished results from chromosome 7 and chromosome 22 indicate it should be possible to sequence into the telomeric repeats at least for some chromosomes.

Session VI ANNOTATION
Chair: Graham Cameron

The session reviewed annotation practice by the high throughput genome sequencing centres and considered a minimum level of annotation that should be provided with sequence submitted by centres to the sequence databases.

From a comparison of the current entries in the databases it was evident that different centres were annotating features in different styles and depth. Annotated features included: neighbouring clones, CpG islands, repeat content, difficulties in obtaining sequence of high confidence, STS content, exons and ESTs (information placed in CDS or miscellaneous category), primary transcripts, mRNA, BLAST data, promoters.

Annotation could be divided into two categories:

1. Book-keeping information i.e. where and how the sequence was obtained.
   This information could only be provided by submitters.

2. Biological annotation of the sequence.

BOOK-KEEPING INFORMATION

It was agreed that a minimum level of annotation consisting of book-keeping information should be required with submissions of finished sequence.
Required features should include:

- clone identification
- source library information
- contact details of sequencing organisation.

Further candidates for required information included:

- detailed location information on neighbouring clones and where mapping information relating to the sequence could be found
- information on the accuracy of the sequence.

It was not clear cut whether data on the accuracy of the sequence should be required. The use of the data to the community would be highly dependent on the way such information was represented. (Suitable methods for conveying accuracy information had been discussed in session II.)

It was suggested that the sequence database providers circulate a proposal regarding required annotation, and that Elbert Branscomb circulate a proposal on the representation of accuracy information.

BIOLOGICAL ANNOTATION

It was agreed that annotation was essential to exploit the sequence information being generated. The centres did not reach a consensus view as to how this should be achieved. In particular, there were differing views on whether annotation should be applied by the sequencers or the databases. However there was agreement that sequence made available to the community should be accompanied by consistent annotation.

Almost all annotation by the participants is by computational methods. These methods assign features and functions on the basis of analogies to existing biological knowledge, either as represented in existing sequence entries, or in databases of sequence patterns characteristic of particular biological functions. This raised two major discussion themes:

- Minimalist vs. Rich Annotation
- Documentation of the evidence for features

Minimalist vs. Rich Annotation

Some centres took the view that the computational methods were either (a) readily available, making it unnecessary for sequencing centres to supply the annotation, or (b) more likely to be applied expertly and consistently by information centres such as the EBI or NCBI. Proponents of this view formed the "minimalist-annotation" camp. Such centres espoused the view that, by and large, their responsibility extends only to annotating with book-keeping information.

Others stressed the need for expert interpretation and refinement of predictions from automatic methods, and also pointed out work on real biological investigation of the function of sequences. They saw it as part of their responsibility to provide "rich-annotation" of their sequence. It was stressed by some participants that whatever standard was applied, consistency of practice would be extremely helpful to the user community.

The particular emphasis of centres reflected the ability to obtain resources for annotation. Some funding agencies indicated that they regarded detailed annotation and development of algorithms as a research exercise and distinct from sequence production. It was therefore an activity in which the genome centres could participate by competing for specific research grants.

Proponents of rich-annotation stressed:

- the role of annotation as a final check of sequence accuracy: e.g., disruptions in exon connections can highlight any sequence errors.
- the importance of expertise, often present within sequencing centres, in interpreting the results from sequence analysis programs.

Proponents of minimalist-annotation stressed:

- the importance of exploiting the expertise of groups not involved in sequencing, but with a biological interest in a particular region of the genome, to add annotation later
- the imprecision of the automatic methods which sequencing centres might use, particularly when applied to complex genomes
- the fact that predictions based on comparison with existing databases would go out-of-date as the databases were updated, necessitating their constant curation
- the difficulty in achieving any kind of consistency between centres.

Documentation of the Evidence for Features

Feature annotation which is assigned on the basis of computational evidence is often presented without any indication of the nature of the evidence for a given assertion. Even where an attempt is made to document the evidence, there is little consistency about how this is done.

A model for detailed machine-parsable representation of computational evidence used in the *Drosophila* database (FlyBase) was presented. This database documents the precise database matches on which genes and coding sequences are annotated. Whilst there was support for better documentation of computational evidence in principle, there were concerns about the practical difficulty of ensuring consistent application of such a detailed standard, and about the need to curate such information.

THE ROLE OF THE SEQUENCE DATABASES

Sequencing centres and funding agencies which adopt a minimalist annotation policy saw a role for the sequence databases in applying computed annotation to the data. There was some support for the concept of an "annotation-engine" which would automatically apply up-to-date annotation to all data, while some doubted the feasibility of such a concept. Another view of the central databases was as a repository for annotation received from the sequencers, with little active application of new annotation. Broadly speaking, the USA favoured a more active role for the databases than Europe.

Some centres felt a clear responsibility for the entries they submitted to the public databases, and wished to see them presented with precisely the annotation and updates they had provided. Any annotation applied by the databases or third parties should be clearly distinguished from the submitted annotation and not mixed with it. Other centres saw annotation as a responsibility of the databases.

It was recognised that there was a need for specialist databases which would provide more detailed annotation than the central databases, and it was agreed that cross-links to such databases were desirable.

GENE IDENTIFIERS

There was a need to agree a format for identifiers for genes whose existence was predicted from sequence analysis, but for which there was no confirmatory experimental evidence. A unique identifier in the form of clone_name.chronological_number was proposed. This would be replaced when the existence of the gene was confirmed and a HGM name allocated. A similar format had functioned well for *C.elegans* and had the advantage of identifying the source clone. There was a need to avoid the situation that had occurred with yeast where the genes had to be renamed. An alternative view was that more information could be provided in the name so that it could be used in the longer term. The HGM nomenclature committee should be consulted for their views on the matter. There was greater support for keeping the name as simple as possible in the first instance.

Session VII NEW TECHNOLOGY
Chair: Rick Wilson

SEQUENCER TECHNOLOGY

Slab Gel Technology

Both ABI and Washington University had modified ABI 377 sequencers to increase the number of lanes to 96 per gel. This had been achieved by increasing the pixel density per scan. Results were promising as tracking did not appear to be a problem, difficulties associated with the additional lanes were at the gel pouring and loading stages. Washington University's modification was being exported to other centres. It was hoped, although there was no hard evidence, that increasing the number of lanes per gel would result in cost savings.

Lloyd Smith at the University of Wisconsin was developing a low cost sequencer with the primary aim of producing long read lengths rather than high throughput. 18 hour run times had produced read lengths of about 1 Kb. The cost of the sequencer would be low, in the region of $29,000; two production machines were now being tested. It was noted that the Licor machines also achieve read lengths of 1 Kb but their major disadvantage was only 1 & 2 colour detection.

The question of patent protection on existing technology was felt to be a possible future issue of contention with patent holders but should not discourage attempts to improve technology.

Capillary sequencers

Lawrence Berkeley was testing a 96 channel capillary sequencer utilising 4% Linear Polyacrylamide (LPA) as the sieving medium. 23 plates of production sequencing had been performed. Currently, reads up to 450 bp was being obtained but at ~450 bp the drop in sequence quality was steeper than with ABI 377s. Problems associated with the reuse and reliability of LPA and capillaries were also in the process of being resolved. Comparison with Molecular Dynamics capillary sequencers were planned.

Molecular Dynamics capillary sequencers had been trialed at Washington University and TIGR. Reads of 400-450 bp were being obtained. It was thought that longer reads would be obtained if the sieving medium was changed from hydroxycellulose to LPA, the change would be dependent on a licence being obtained from the patent holders, Beckman. At the present time similar levels of sequencing reagents were being used as for conventional slab gels.

Microchannel System

The JGI were in the process of developing a microchannel system. Microchannels were etched onto a plate and into these the sieving medium, hydroxycellulose, was pumped. Further automated of the system was planned with the development of a gel loader. Read lengths of around 500 bp were being obtained with as little as 1/100 of the sample required for ABI 377s. Currently a single plate with 96 channels could be scanned but the potential scanning capacity was 4 plates at once i.e. 384 samples. The 96 channel machine was 1 year away from production and the 384 version, 2 years.

AUTOMATION

The DOE had recently funded several groups to develop genome sequencing technologies. Amongst these the Whitehead Institute had been funded to develop robots with commercial potential.

Several groups reported working on automation of various procedures The JGI was developing its automation around Hydrahead 96 channel pipettes. Machines for handling frozen stocks and setting up PCR reactions were now working. The next target was setting up sequencing reactions. At UTSW the problems associated with the automated system had been in achieving PCR reliability in rather than with the robots themselves. Automation at the Sanger Centre had concentrated on a system for rearraying finishing templates.

CHEMISTRY/BIOLOGY

1. **Enzymes**
   IMB at Jena, Amersham and Perkin-Elmer were developing mutant polymerases to read through problem sequence. It appeared that if mutants could read through difficult sequence they were less effective than current enzymes on non-problem sequence.

2. **Dyes**
   Richard Gibbs reported that he was using his BODIPY dyes exclusively but there would be patent repercussions from ABI if these were made available commercially. BigDyes had been shown to work on templates of 2.5 Mb making reads possible directly from microbial DNA and YACs.

3. **Vectors**
   It was reported that Bob Weiss at Utah had a plasmid vector system which could take 12-15 Kb inserts. Initial data suggested that they were stable. Other cloning vectors such as circular YACs and BACs containing yeast centromeres and HIS markers had been discussed in session V.

4. **Shatter Libraries**
   Washington University and the Sanger Centre had developed shatter libraries to read through sequence gaps. A PCR product or restriction enzyme fragment was sonicated to generate 100-500 bp fragments which were cloned into M13 or pUC. In GC rich regions reads were possible using 100 bp clones, as the short clones prevented secondary structure forming and interfering with enzyme processivity.

5. **Chemical Sequencing**
   Chemical sequencing had been revisited by some groups for problem regions but had not been worth pursuing.

NOVEL TECHNOLOGIES

The groups had very little information on areas other than improvements on existing technology. Most novel developments were felt to be too distant and therefore it was necessary to concentrate on adjusting current technologies in order to achieve the increase in sequence production in the short term. In order to facilitate technology development sequencers should be encouraged to liaise with each other and with outside technology developers. One reported development was in detection systems. A mass spectroscopy detection system was being developed which could increase the number of labels from 4/5 to 500 enabling 100 reactions to be analysed in parallel.

Session VIII MOUSE GENOME SEQUENCING & FUTURE MEETINGS
Chair: Francis Collins

Francis Collins chaired this session in the absence of Michael Morgan. A significant number of participants had to leave before the end of the session and before the votes were taken.

DATA RELEASE POLICY FOR ORGANISMS OTHER THAN THE HUMAN

In addition to human genome sequencing several groups are starting to produce mouse genomic sequence. The NHGRI was encouraging grantees to devote up to 10% of their effort to sequencing the mouse genome. It was therefore an appropriate time to decide on the policy for release of mouse sequence data. All present were in unanimous agreement that the policy of immediate data release, as agreed for human genomic sequence, should be applied to the mouse.

It was proposed that this policy should be extended to high throughput sequencing projects for all organisms. It was agreed that the participants were not in a position to make policy for projects outside of their own area, but should aim to influence policy by stressing the value of early data release for research and for co-ordination of the sequencing effort. Several members of the group felt in a difficult position owing to national funding agencies not endorsing early data release. It was stressed that any statement on this issue should reflect the personal views of the participants based on their experiences of participation in the human genome sequencing project.

It was proposed by Francis Collins that a statement to this effect should be issued following approval by all participants. The proposal was seconded by Ari Patrinos and supported by all present.

The following statement was circulated and approved by the participants.
"As extensive determination of the genomic DNA sequence of several organisms proceeds, it is increasingly clear that this information has enormous and immediate scientific value, even prior to its final assembly and completion. Delaying the release of either unfinished or finished genomic DNA sequence data serves no useful purpose and actually has the effect of slowing the progress of research. Therefore, the attendees at the Third International Strategy Meeting on Human Genome Sequencing agreed unanimously to support, as individual scientists, the view that all publicly-funded large scale DNA sequencing projects, regardless of the organism, should deposit data immediately into the public domain, following the same guidelines that have previously been adopted by this group for human genomic sequence (http://hugo.gdb.org/bermuda.htm). We will continue to adhere to these principles and urge all other scientists and policy-making groups involved in large scale sequencing to adopt them as well."

FUTURE MEETINGS

It was agreed that the strategy meetings were essential for the co-ordination of human genome sequencing. The current number of attendees (about 50) allowed important issues to be debated fully and therefore should not be increased, although if additional groups started to contribute to the project they should be invited.

Several participants expressed the view that difficulties in reaching the current venue could discourage attendance at the meetings. To encourage all invitees to attend, a model of alternating venues within reach of Dulles and Heathrow airports was proposed.

The alternatives were put to the vote.

7 for remaining in Bermuda
19 for moving the meeting
22 either abstained or were absent for the vote

The exact venue for the next year would be decided at a later date. The date of the next meeting would probably be put back a month or two as HGM '99 would probably be taking place at the end of February in Australia. To aid discussions more formal presentations on particular areas and reports from working groups would be organised, the programme would be co-ordinated by the Wellcome Trust, NHGRI and DOE.

APPENDICES

Appendix 1 Reports from Genome Sequencing Centres

Appendix 2a Summary of the Second NHGRI Quality Assessment exercise

Appendix 2b Summary of the Sanger Centre Internal Quality Assessment exercise

APPENDIX 1

Participants at the meeting were asked to provide information on progress made and the amount of sequence produced prior to the meeting.

Submissions have been updated following discussions at the meeting and annotated with data on the predicted scale-up, and sequencing and mapping strategies.

The date of submission of the progress report is indicated after the participant's name

STANFORD UNIVERSITY

R MYERS

25 Feb 1998

* total human bp finished	0.65 Mb
* total bp submitted to GenBank	0.51 Mb
* why the difference between finished & submitted?	In annotation
* total finished this past year	0.35 Mb
* unfinished	3.5 Mb (in progress)
* largest submitted contig length	152 Kb
* some sort of quantitative measure of quality	error rate 1 in 65 Kb (PHRAP analysis)
* critical issues are for scaling up	Funding

* Predicated scale-up	current rate	7 - 8 Mb/year
	April 98-April 99	12 Mb/year
		22 Mb/year
		30 Mb/year

Strategy
It was reported that in the past year, after the first NHGRI QA exercise, the focus at Stanford had been on generating high quality sequence. The mapping strategy was concentrating on ordering STSs on a high resolution Radiation Hybrid map prior to screening Pieter de Jong's BAC library. The rationale for this strategy was that well distributed STSs should result in more even coverage and more contiguity than randomly distributed STSs. The optimal STS density was being investigated.

UNIVERSITY OF TOKYO

Y SAKAKI

20 Feb 1998

* total bp finished	5.1 Mb (submitted to JST)
* total bp submitted to GenBank	2.9 Mb
* why the difference between finished & submitted?	to be submitted soon
* total finished this past year	2.4 Mb
* unfinished	5 Mb (in DDJB)
* average submitted contig length	1 Mb (minimum required by JST)

* some sort of quantitative measure of quality

  more than 99.9% (compared to data in public databases)

* critical issues are for scaling up

  It was reported that the new genome centre funded by STA (Science Technology Agency) would be opened in October 1998 and this would increase the available space and funding

* Predicted scale-up	October 1998-99	8 Mb/year
	October 1999-2000	15 Mb/year
	October 2000 onwards	30 Mb/year

Strategy
The current sequencing strategy has focused on nested deletions with a 3-4 fold redundancy. It would not be possible to simply scale-up this approach much further and therefore there would be a move towards a new strategy combining both nested deletion and shotgun sequencing steps.

UNIVERSITY OF WASHINGTON

P GREEN

25 Feb 1998

* cumulative total bp finished	3.65 Mb (human, non-redundant)
* total bp submitted to GenBank	3.65 Mb
* why the difference between finished & submitted?	NA
* total finished this past year	3.06 Mb
* unfinished	under 1 Mb
* largest contig	1.7 Mb
* average submitted contig length	640 Kb

* some sort of quantitative measure of quality

In the chromosome 7 project we have sequenced 240,674 bp twice because of clone overlaps. There were four discrepancies found, 3 of which were due to cosmid mutations (a single base indel in a poly A run in each case) and one was a sequencing error. So our best estimate of our error rate is one in 2 x 240,674, or about 1 per 480 Kb.

No errors were found in any of our sequences in the NIH quality checking exercise.

* critical issues are for scaling up

We cannot scale-up within the university due to inadequacies of space and of academic career incentives for the key personnel, and are forming a company to do genomic sequencing outside the university. The most critical issue for us is whether the NIH will back this kind of effort, which is unclear.

It was reported at the meeting that the NIH would provide grants to companies if the companies could meet the conditions of funding and were competitive.

* Predicted scale-up	July 1998-99	10 Mb/year 100 Mb over 3 years

Strategy

The strategy was directed towards achieving maximum sequence contiguity. Currently the mapping strategy involved obtaining YACs at a 2 fold depth: the YACs were subcloned into cosmids and an MCD map generated. The MCD map was important to check the assembly, check clone validity and to identify the most efficient tiling path. In the near future there would be a move from YACs to BACs. There would be some direct sequencing from BACs, but as it was difficult to interpret MCD maps generated directly from BACs some subcloning into cosmids would be continued.

It was noted that centres focusing on contiguity would necessarily generate a smaller proportion of unique sequence relative to the total sequence output to owing to the greater number of overlaps that would be sequenced.

There had been no region of the sequence that once cloned appeared to be unsequencable.

ACGT UNIVERSITY OF OKLAHOMA

B ROE

26 Feb 1998

* total bp finished	4.13 Mb
* total bp submitted to GenBank	4.13 Mb
* why the difference between finished & submitted?	NA
* unfinished	2.8 Mb
* average submitted contig length	220 Kb
* largest contig	1.45 Mb (Di George region)

* some sort of quantitative measure of quality

> Sequence data revisited after QA exercise, used PHRED and PHRAP to ensure data quality met the agreed standards

* critical issues are for scaling up -

> The University meeting its commitment to provide extra space.
> Scalability of project centered management.

Predicted scale-up	currently	3 Mb /year
		7 Mb/year
		10 Mb/year
		15 Mb/year
	achieved by 2004	40 Mb/year

Strategy
To date the SRM has been obtained from collaborators. Depending on funding, the capability to perform restriction analysis for clone validation would be developed in house. However, there were no plans to develop any further mapping capability and it was hoped that the mapping groups would continue to provide clones. A strategy of primer walking for prefinishing, similar to that at UT Southwestern, was being used.

DOE JOINT GENOME INSTITUTE

A. CARRANO

22 Feb 98 and post meeting

* total bp finished 7.67 Mb

Total bases that have gone through assembly into a single contig but have not been validated, annotated, or submitted approx. 2Mb

* total bp submitted to GenBank 5.67 Mb

* why the difference between finished & submitted?

For those sequences not in the pipeline, the difference is due to:

1) verification of sequence
2) analysis of the sequence
3) annotation

* unfinished 10 Mb

* average submitted contig length 50.2 Kb (mean - total sequence/number of clones)

* largest contig 1.02 Mb

* some sort of quantitative measure of quality

By summing the error rate for each base in the submitted clone:
For a total of 1,250,069 bases submitted (22 clones), the error rate is calculated to be 1.53 per 10,000 bases. (max. estimate)

* critical issues are for scaling up

a) staffing and training (80 hires in first year of operation)
b) scaling up 10X in throughput in one year
c) operating at three different sites as a virtual center while renovating and moving into a new commercial facility at Walnut Creek, CA. in the summer (at least 80% of DOE sequencing).
d) managing three interim infrastructures and transitioning different sequencing approaches into a one process (shotgun and some directed closure) and infrastructure
e) keeping costs as low as possible in a start-up year

* Predicted scale-up	October 1997-98	20 Mb/year
		40 Mb/year
	in 2000	100 Mb/year

Strategy

The generation of SRM was not limiting, currently 168 Mb was available in mapped clones with an aim to maintain a clone supply 6-12 months ahead of sequencing. In the past 6 months 90% of clones going into the sequencing pipeline were in contigs greater than 1 Mb.

UT SOUTHWESTERN

G. EVANS

25 Feb 1998

* total human bp finished	4.35 Mb
* total bp finished submitted to GenBank	4.35 Mb
* why the difference between finished & submitted?	NA
* unfinished	6.5-7 Mb
* average submitted contig length	~200 Kb
* largest submitted contig length	807 Kb

* some sort of quantitative measure of quality

Better than 1 in 10,000 bp (early cosmid data revisited after QA exercise)

* critical issues are for scaling up
continued funding
personnel

* Predicted scale-up	current rate	12 Mb/year
	August 1998-1999	24 Mb/year
		48 Mb/year

Strategy

It was reported that clone production was exceeding the sequencing capacity at the current time. As for many groups the major bottleneck in the process was finishing. Several Mb of unfinished sequence had built up because of the presence of repeat structures which required specialist finishing to complete. A semi-automated prefinishing stage consisting of primer walking was being used. The strategy was cost effective because of the low cost oligonucleotides made on a MERMAID system (reagent cost 10-12 cents/base). Sequence was scanned by PRIMO for low PHRED and PHRAP scores and primers chosen accordingly. This step significantly reduced the number of gaps that needed closing. In the future, primer walking would be directly from the BAC target clone in order to avoid the potential for subclone errors being incorporated into the finished sequence.

ACGT - ABI

E CHEN

24 Feb 1998

* total bp finished	5.0 Mb
* total bp submitted to GenBank	5.0 Mb
* why the difference between finished & submitted?	NA
* total finished this past year	2.6 Mb
* average submitted contig length	200 Kb
* largest submitted contig length	700 Kb

* some sort of quantitative measure of quality

> We did not participate in the recent QA exercise. We routinely check our data quality by PHRED scoring test. A few selected checking between the overlapping regions indicated the error rate is less than 1 base per 10,000 bp.

* critical issues are for scaling up

> Building up infrastructure is the hardest part for us. I believe the most ideal rate is to go by 2X increment each year.
>
> Realistic expectation: throughput at about 10 Mb/year and cost at about $0.40/base including mapping.

It was reported that there was currently a gap of funding for human genome sequencing for one year. Commercial projects would be taken to fill the gap. In common with many groups the Centre had seen a high level of personnel turnover in recent years.

Strategy

The Centre has been reliant on sequence-ready clones supplied by collaborators. Experience had shown that, on sequencing, there were a significant number of instances where results obtained did not agree with the sequence-ready map. To address this a capability to validate sequence-ready maps prior to sequencing was required.

IMB - JENA

A ROSENTHAL

26 Feb 1998

* total bp finished	5.29 Mb
* total finished this past year	3.79 Mb
* unfinished	7 Mb
* average submitted contig length	280 Kb
* largest submitted contig length	1 Mb

* some sort of quantitative measure of quality

    1 in 70-80 Kb (from overlaps).
    Cross checking data within German consortium, not resequencing

* critical issues are for scaling up
    Dependent on funding from the BMBF for human sequencing, current funding up to the year 2000 although there will be a second funding period
    Insufficient space: space being rented outside the institute
    Most of the SRM had been obtained from collaborators, in order to scale-up there will need to be a greater in house mapping capability

* Predicted Scale-up	January 1998-1999	9 Mb 15 Mb

Other members of German sequencing consortium had produced 600 Kb of sequence with targets of 1 and then 2 Mb for the following years

The amount of human genome sequence that would be produced by IMB would also depend on the amount of effort directed towards other genomes, including the mouse and *Dictyostelium*.

BAYLOR COLLEGE OF MEDICINE

R GIBBS

22 Feb 1998

* total bp finished	6.5 Mb
* total bp submitted to GenBank	5.9 Mb

* why the difference between finished & submitted?

Difference between progress and submitted is mostly due to clones being "in progress", as a relatively small amount is stuck in the pipeline. Finishing reactions are the bottleneck, but not by much.

* total finished this past year	3.5 Mb
* unfinished	2.3 Mb (closure) 3 Mb (random)
* average submitted contig length	130 Kb
* largest submitted contig length	650 Kb

* some sort of quantitative measure of quality

> Quality is better than 1 in 10,000 as per the recent exercise. We had one clone with a putative "missassembly" but this was a 4.5 year old fragment that was done by a grad student and the clone was lost and regrown so we think it rearranged etc. Therefore we claim a consistent rate of better than 1/10,000.

* critical issues are for scaling up

> Ongoing issues are staff and institutional support and Senior Personnel dedication.

* Predicted scale-up	current	7.5 Mb/year
	July 1998-99	30 Mb/year

Strategy

The SRM was being obtained from collaborators, in addition to some in house mapping. The SRMs obtained from collaborators were generally of high quality but all maps were verified by restriction mapping prior and post sequencing. Last year's sequencing target had not been met because the ramp-up of production had not occurred early enough. The centre was concentrating on producing on contiguous sequence.

THE INSTITUTE FOR GENOMIC RESEARCH (TIGR)

M ADAMS

23 Feb 1998

* total bp finished	6.4 Mb
* total bp submitted to GenBank	6.4 Mb
* why the difference between finished & submitted?	NA
* total finished this past year	3.7 Mb
* unfinished	2.7 Mb (closure) 2.5 Mb (random)
* average submitted contig length	164.7 Kb
* largest contig	500 Kb

* some sort of quantitative measure of quality

quantitative quality measure - one gap closed, and no base pair errors in 200 Kb.

* critical issues are for scaling up

1. Technology. Better technology required to reduce the number of sequencers, technicians and space required per Mb to be able to achieve 10x scale-up. 96-lanes is an excellent start, since it doubles lane density per linear bench space/equipment cost.

2. Space. At maximum capacity, but a new building planned to double the space. Funding not yet completely secured for the new building.

3. Cost is not critical as current costs within a factor of 2-3 of targets based on NHGRI projections.

4. Management. Throughput will depend to a large extent on what the management structure will support, as long as sequencing is personnel-intensive (and closure will always be personnel-intensive). It would be unlikely to ever have 100 people devoted to human sequencing at TIGR.

* Predicted scale-up	next year 13 Mb doubling from then on

Strategy

It was reported that the SRM had been generated using a 'just in time mapping' strategy consisting of screening and end sequencing. The contig figures above had therefore been generated without directed effort towards obtaining long contigs. There would now be more directed effort to obtain greater contiguity.

Random BAC end sequencing is being done on NIH-approved libraries from Shizuya and de Jong to facilitate selection of BAC clones for complete sequencing. BAC end sequences are available from GenBank, a WWW interface at TIGR (for searching completed BACs to look for overlapping BACs based on end-sequence matches), and by ftp (ftp.tigr.org).

Total BAC end sequences (as of 3/1/98): 56,733
Total expected by 10/98: 300,000 (with Lee Hood's lab)
Total expected by 10/99: 600,000 (with Lee Hood's lab)

WHITEHEAD INSTITUTE

E LANDER

25 Feb 1998

* total bp finished 7.4 Mb * (inc. 0.4 Mb mouse)

* total bp submitted to GenBank 7.0 Mb*

* why the difference between finished & submitted?

  The rest are in the "sign-off process" before things are allowed out the door i.e. restriction map checking

* total finished this past year 6.9 Mb

* unfinished 13.3 Mb

* average submitted contig length Typical BAC size

* largest submitted contig length ~ 320 Kb

* some sort of quantitative measure of quality

  Quality measure: 10 errors in 200 Kb were reported in the QA exercise.

  [3 were bookkeeping errors -- correct in our finishers files but incorrect in the file transmitted to GenBank; 2 are reported "errors" with which we aren't sure we agree; and 5 are clear errors (including 2 instances of deletions within an M13, causing a local deletion)].

* critical issues are for scaling up

  Increasing the efficiency of processes and a systematic mechanism for closing gaps

* Predicted scale-up currently 7-7.4 Mb/year
  next Bermuda year 14 Mb
  next financial year (July1-June30) 16 Mb

Strategy

It was reported that a strategy relaying on STSs screening to generate the SRM had been used. This had given sufficient clones but only in relatively small contigs. To increase contiguity, the critical issue was generating more markers in a targeted manner. Probes were being generated from YACs to close the map Mb by Mb. Previous experience from mouse indicated that chimeric YACs were not a problem as they were detectable where four-fold depth of coverage was obtained.

There had been increases in efficiency and automation: short sequencing reads been extended to long (10 hour) runs, shotgun and prefinishing processes had been automated. Each automated process (station) was designed to be as stand alone as possible, a fully integrated system was felt to be unnecessary and would result in a loss of flexibility. In the finishing process there was a need to increase automation but more importantly a systematic way of tackling sequencing and cloning gaps should be found.

WASHINGTON UNIVERSITY

B WATERSTON

25 Feb 1998

* total bp finished	24.5 Mb
* total bp submitted to GenBank	22.6 Mb

* why the difference between finished & submitted?

awaiting analysis and annotation

* total finished this past year	19.7 Mb
* unfinished	37 Mb (shotgun finished) 9 Mb (assembled) 10 Mb (in shotgun, not assembled)
* average submitted contig length	175 Kb
* largest submitted contig length	967 Kb

* some sort of quantitative measure of quality

The recent checking exercise found 6 errors in 200,000 base pairs of our human genomic sequence data. We routinely run the same type of check on ALL finished data before it is submitted to GenBank. We do not submit sequences with gaps or N's. Previous checks of *C. elegans* and *S. cerevisiae* genomic sequence data generated by our laboratory indicated a similar low error rate.

* critical issues are for scaling up

Space will be an issue at some point. Recruitment and retention of experienced senior staff is a crucial issue. As far as the actual business of sequencing, we need better biological and software tools to speed finishing. Given current technology, we believe that we can scale-up to a level of finishing 200 Mb per year within three years, given adequate funding.

* Predicted scale-up	July 1998-99	60 Mb
		100 Mb
		120-150 Mb

Strategy

The sequence-ready clone output was now in the range of 2 Mb/month which was sufficient for the current sequencing capacity. A strategy of using STSs to generate the SRM had resulted in a lack of long range contiguity. In order to increase contiguity of the map, additional probes were being generated from YACs to fill the gaps.

The human genome sequencing capacity is set to increase with the completion of the *C.elegans* genome in 1998.

SANGER CENTRE

J ROGERS

25 Feb 1998

* total human bp finished	36.02 Mb
* total bp submitted to EBI	34.85 Mb
* why the difference between finished & submitted?	awaiting analysis and annotation
* total finished this past year	21.4 Mb
* unfinished	28 Mb
* average submitted contig length	one BAC/PAC clone (100-150 Kb)
* largest submitted contig length	962 Kb

* some sort of quantitative measure of quality

> We weren't part of the US checking exercise, but in an exercise recently carried out in the Sanger Centre across all organisms we found 4 errors in 1.1 Mb.

* critical issues are for scaling up

> We have space, we have some money, although how much we shall be able to achieve is still somewhat in question. Extra funding requested to meet quality and quantity targets. The biggest problem we probably face at present is recruitment and retention of good staff, especially finishers.

* Predicted scale-up	40 Mb/year 80 Mb/year 100 Mb/year

Strategy
An STS map at a density of 15 markers/Mb was being used to generate the SRM. 280 Mb of clones in contigs were available, and were being fed into the shotgun at a rate of 1.5 Mb/month. There had been some directed closure on chromosome 22 using targeted markers from YACs. Currently there were 10 gaps in the Sanger Centre region of chromosome 22 but untapped resources were still available to try to close them and the sequence would be completed by the end of the year.

Improvements were being made to the finishing process: automating the system for rearraying the finishing templates, a new implementation of FINISH to chose the finishing reads, and improved autoediting to produce the consensus sequence. These changes would allow finishers to concentrate on problem solving. In the preceding year many gaps had been closed using shatter libraries and the introduction of ‘black tags’ to flag sequence of lower confidence.

The human genome sequencing capacity is set to increase with the completion of the *C.elegans* genome in 1998.

Appendix 2a

2nd NHGRI Large-Scale Sequencing QA Exercise

METHOD

- Selected four finished clones, at random, totaling 200 kb, from each participating sequencing group (all NHGRI human plus *D. melanogaster*)
- Data checked was selected from that deposited as 'finished' as of September, 1997
- Assigned each set of four clones to two checkers chosen from among the participants; groups exchanged data files and bacterial isolates/DNA
- Checkers re-assembled files and analyzed data. If error rate was better than 1 in 2000, resolved discrepancies by further analysis (resequencing).
- Each group responded to checker's reports
- Most groups checked assembly by restriction analysis

Total number of clones available for checking *as of 9/97*: 420
Total number of clones selected for exercise: 37 (a total of 1.7 Mb tested)

2nd NHGRI Large-Scale Sequencing QA Exercise

RESULTS

Single-base discrepancies–number of clones at[a]:

<1/10000	1/10000-1/5000	1/5000-1/2000	>1/2000	Total
22	10[b]	1	3	36*

[a] These numbers are based on the results that indicated the higher error rate among the two reports, for each individual clone; these numbers do not take into account the producer's responses.

[b] For 7 out of the 10 clones in this category, one of the two checkers actually evaluated those clones as having fewer than 1 in 10000 errors.

- Total number of single-base discrepancies (conservative aggregate of two checkers): 230/1.7 Mb. Total excluding the clones worse than 1 in 2000: 120/1.59 Mb

- About 2/3rds (133) of the single-base discrepancies were substitutions, 1/3rd (73) were insertions or deletions, based on 206 cases of single-base errors where precise information was provided

Other errors (not exclusive of single-base errors)

1 mis-assembly, origin unknown
1 possible mis-assembly (1900-base deletion); may be a clone instability
1 clear clone instability (~250 bp deletion)
1 likely clone instability (~650 bp deletion)
1 annotated gap closed (75 bp)
*1 wrong clone sent (clone tracking error)

2nd NHGRI Large-Scale Sequencing QA Exercise

SUMMARY

- **Caveats:** Variability due to sampling; variability in checking

- **Most groups are sequencing at or very close to standards:** Most groups are at 1 in 10000 or better, summed over all clones. Numbers in the table are conservative and do not include the producer's responses, consideration of which will improve the error rates. However, most of the producers responses agree with the checkers' reports.

- **Good concordance between checkers' reports**: For single-base errors, both checkers agreed on the general quality of the project (according to the bins in Table 1) 28 of 37 times, and were very close in all other cases. In 11 of 19 clones where error type and location appear in the report, there is at least a 50% overlap in the precise identified errors. But there were still some puzzling differences between the identified errors in an individual clone, especially when there were a lot of errors or trace data were considered poor by checker. For other types of error (deletions, etc.), both checkers agreed in all but one case.

- **The exercise reveals useful information about the kinds of error:** Clone instabilities (small deletions) were a small but significant problem—small deletions may be hard to detect with routine protocols. Single-base errors often occur in regions where sequence data quality is good—more than half could be resolved unambiguously by re-editing the original data without need to re-sequence (36/53 errors). (Some of this was confirmed by resequencing).

Appendix 2b

SUMMARY OF CHECKING EXERCISE AT THE SANGER CENTRE, JANUARY 1998

- **17 PROJECTS(~1.1Mb) COVERING 4 DIFFERENT ORGANISMS WHICH HAD BEEN SUBMITTED IN THE LAST 24 MONTHS WERE CHOSEN.**

- **GENERALLY COVERAGE WAS GOOD AT 5 - 7 FOLD.**
  SOME OLDER PROJECTS WERE THINNER AND HAD REQUIRED SUBSTANTIAL AMOUNTS OF ADDITIONAL FINISHING.

- **EDITING ERRORS**
  FOUR ERRORS DETECTED. ALL INCORRECT BASE CALLS.

  ONE WAS MISINTERPRETATION OF WEAK G AFTER A (A PROBLEM INHERENT IN THE USE OF OLDER TERMINATOR CHEMISTRY).

  ANOTHER DUE TO AUTO-EDIT ERROR IN A LOWER QUALITY REGION.
  ACTION 1: AUTO-EDIT NOW PREVENTED FROM EDITING IN A REGION OF TWO READS ONLY.

  ACTION 2: RANDOMLY RECHECK OLDER CLONES FOR POSSIBLE ERRORS DUE TO MIS-CALLING WITH THE OLDER TERMINATOR CHEMISTRY.

- **ANNOTATION TAGS (BLACK TAGS) ARE USED.**
  ACTION: TAG FORMAT NOW STANDARDISED. VERIFIED THAT COMMENTS ARE CARRIED THROUGH INTO THE SUBMITTED ANNOTATION.

- **FURTHER CHECKING WITHIN THE DIFFERENT SANGER TEAMS WILL CONTINUE TO ENSURE CONSISTENCY**

Mark Guyer

CONFIDENTIAL

DRAFT

Report of the Second International Strategy Meeting on Human Genome Sequencing held at the Hamilton Princess Hotel, Bermuda, on 27th February - 2nd March 1997

Summary

- The principles enunciated at the first International Strategy meeting, of rapid data release and public access to the primary genomic sequence, were reaffirmed.
- Scientists and funding agencies should take the necessary steps to ensure that the principles are adhered to by all participating organisations.

Sequence Quality Standards

The following standards were agreed:

- The nucleotide error rate should be 1 error in 10,000 bases or less for most sequence.
- Assemblies should be verified by restriction digest using two or more restriction enzymes.
- Gaps in sequence. The agreed long term goal is no gaps, recognising that this is not yet routine.
- Closing gaps is the responsibility of the original sequencer.

The following proposals were endorsed by the participants:

- It was agreed that a useful trial to assess sequence accuracy would be to perform a data exchange exercise. Raw sequence data would be exchanged among sequencing centres, centres would reassemble the data and identify outright discrepancies or ambiguities with reference to the sequence submitted to the database. These would be resolved by further consultation or resequencing. The same data sets would be sent to two centres which would hopefully engender competition to detect errors.
- All sequence reads should be archived in a retrievable form.
- Sequencing centres should define explicitly how error rates and costs have been calculated.

Sequence Submission and Annotation

Sequence data should be classified simply as "finished" or "unfinished" and should be stored in distinct databases; consideration should be given to establishing a public database for unfinished sequence data.

Sequence annotation should be standardised if possible, and include the following information:

- Error estimation such as PHRED AND PHRAP data.
- Enzymes used to verify assembles, and sizes of fragments produced.
- Exact details on how to assemble adjacent clones, with a minimum of 100 bp of overlapping (preferably unique) sequence between clones for verification.
- Gaps must be sized and the surrounding sequence oriented and ordered. The methods used for sizing, and reasons for not closing the gap should be stated.
- If features such as coding sequence and splice sites are included in the annotation, it should be stated if they were identified experimentally or by computer predictions.
- Unfinished sequence; it should be stated how near the sequence is to completion.

Potential development of a database listing all gaps in 'finished' sequence.

Sequence Claims and Etiquette

Mapping investment does not automatically entitle sequencing claims over the same region until a sequence ready map has been generated.

Potential conflicts with other sequencers to be resolved by early communication.

Collaborations with groups with a biological interest in a region should be subject to the same principles of data release and communication.

Investigate whether the Human Sequence Map Index should be relocated to be more closely associated with the other major human sequence databases.

Claims allowed on the Index:

- Duration - maximum 1 year.
- Size of region - minimum 1 Mb; regions to be defined by Genethon markers if possible, other agreed and available markers if not.
- Maximum amount - in the order of three times the sequence released by the centre in the preceding year.
- Sequence claims must span the entire region between, and including, the delimiting markers.

Next Meeting

- To be held at the end of February 1998 in Bermuda (dates to be confirmed)

Report of the International Strategy Meeting on Human Genome Sequencing held at the Princess Hotel, Southampton, Bermuda, on 25th-28th February 1996

Aims of the Meeting

To discuss mechanisms to co-ordinate, compare and evaluate different strategies for human genome mapping and sequencing.

To consider the potential role of new technologies in sequencing and informatics and to discuss different scenarios for data release.

Summary

The following principles were endorsed by all participants. These included officers from, and scientists supported by, the Wellcome Trust, the UK Medical Research Council, the NIH NCHGR (National Institute of Health, National Center for Human Genome Research, the DOE (U.S. Department of Energy), the German human Genome Programme, the European Commission, HUGO (Human Genome Organisation) and the Human Genome Project of Japan. It was noted that some centres may find it difficult to implement these principles because of legal constraints and it was, therefore, important that funding agencies were urged to foster these policies.

Primary genomic sequence should be in the public domain.

It was agreed that all human genomic sequence information, generated by centres for large-scale human sequencing, should be freely available and in the public domain in order to encourage further research and development and to maximise its benefit to society.

Primary genomic sequence should be rapidly released.

- Sequence assemblies should be released as soon as possible; in some centres, assemblies of greater than 1 kb would be released automatically on a daily basis.

- Finished annotated sequence should be submitted immediately to the public databases.

It was agreed that these principles should apply to all human genomic sequence generated by large-scale sequencing centres, funded for the public good, in order to prevent such centres establishing a privileged position in the exploitation and control of human sequence information.

Co-ordination

In order to promote co-ordination of activities, it was agreed that large-scale sequencing centres should inform HUGO of their intention to sequence particular regions of the genome. HUGO would present this information on their World Wide Web page and direct users to the Web pages of individual centres for more detailed information regarding the status of sequencing in specific regions. This mechanism should enable centres to declare their intentions in a general framework whilst also allowing more detailed interrogation at the local level.

SESSION I - INTRODUCTION

CHAIR: Jim Watson

In his introduction, Jim Watson spoke of his hopes that the sequencing of the human genome would be completed in his lifetime. He urged the community to proceed in a spirit of friendly competition to reduce costs but in a co-ordinated manner to minimise duplication and redundancy. He considered that the truly competitive phase should be in utilising the primary sequence data to identify disease genes.

This session then comprised brief strategy statements from representatives of most of the sequencing centres represented at the meeting.

Washington University, St Louis and The Sanger Centre, Hinxton

Bob Waterston spoke on behalf of both centres and expressed his hopes that the co-operation and openness that had developed with the *C.elegans* sequencing programme could be continued into the human programme.

The strategy adopted by the two centres was to use an STS(Sequence-Tagged-Site)-based approach to construct a minimal tiling-path from large-insert bacterial clone libraries. Sequencing then involved a mixture of shotgun and directed approaches to produce finished sequence which would be annotated and submitted to the public databases. Maps and assembled sequence data would be made available *via* ftp to facilitate co-ordination and encourage further research by the wider scientific community. The aim was to produce sequence data with a high level of contiguity and accuracy (99.99% in most places). The two centres would be targeting efforts on chromosomes 22, X, 7 and 6. In general, a systematic approach would be adopted but directed sequencing may be used for some areas of importance in disease.

Whitehead Institute/MIT Centre for Genomic Research

Eric Lander proposed a similar mapping strategy using BACs (Bacterial Artificial Chromosomes) anchored by STSs and an M13 shotgun sequencing strategy with directed closure. The distinctive focus of the Whitehead Institute's approach would be the high level of automation proposed throughout the process; the Sequatron. The Whitehead had already developed an automated system for the selection of BACs using STSs and their characterisation by fingerprinting. A hands-off assembly process was also planned using informatics-directed closure. The Centre's goal was to produce 5Mb of sequence in the first year.

The Institute for Genomics Research

Craig Venter expressed concern that the estimated cost of developing sequence-ready maps by conventional approaches was likely to be of the order of $100m and that the maps produced were unlikley to be complete. He proposed a different approach which involved sequencing the ends of BAC clones to produce sequence tags every 5kb. He anticipated that this would reduce the cost of producing a sequence-ready map for the whole genome to $10m and would enhance international collaboration by providing unique, uniformly spaced tags across the genome.

University of Washington, Seattle

Lee Hood endorsed the use of BACs for developing sequence-ready maps. He proposed a similar strategy to that of the previous speakers and was planning to collaborate with Craig Venter using the end-sequence approach to produce Sequence-Tagged-Connectors (STCs) every 4kb;assuming a 20-fold coverage of BACs with 150Kb inserts (400,000 BACs). He recognised the difficulties in resolving repetitive elements using this approach particularly long elements (LINEs) which comprised approximately 11% of the genome. However, the approach should be very amenable to automation since it only required two processes; isolating DNA and sequencing DNA. He endorsed the general principle that data should be rapidly released and of high quality. The software that Phil Green had developed at Seattle to assess quality and eliminate manual editing should address both these issues.

Japanese Human Genome Project

Naotake Ogasawara explained that the Japanese Human Genome Project was funded by the Science and Technology Agency, the Ministry of Education, Science, Sports and Culture and the Ministry of Health and Welfare; the project heads would be Ken-ichi Matsubara and Yoshiyuki Sakaki. Japan also had a major interest in the rice genome project. Human sequencing focused on regions on chromosomes 16 (HLA), 14 (IgH), 4 (TCR), 21 (Down's critical region), 22 (IgL). P1 and cosmid maps were being developed for chromosomes 21, 17, 11, 8, 6 and 3. The two key technologies were the use of flow-sorted chromosomes to make chromosome-specific cosmid libraries and the development of a nested-deletion sequencing method. With current pilot funding, it was hoped that 1-2Mb of sequence would be available in the public domain by the end of the year with a goal of 5Mb for the first year of funding (October 95 - October 1996).

Lawrence Livermore National Laboratory

Tony Carrano stated that the primary human genome sequencing target for LLNL would be chromosome 19; a cosmid map of the 50Mb chromosome had been published in December in Nature Genetics. A shotgun strategy with directed closure was proposed for the chromosome which had a high proportion of Alu repeats. The Laboratory would provide clones and map information to other groups. Comparative mouse studies would be done in collaboration with Rick Woychik and Lisa Stubbs at Oak Ridge National Laboratories. The LNLL was also planning to use ESTs available through the IMAGE consortium to pull out BACs in a genome-wide, high throughput hybridisation strategy.

Baylor College of Medicine

Richard Gibbs proposed a similar strategy to previous speakers, focusing on chromosomes 12 (12p1.3) and X (Xq28 and Xp22). The strategy would include some reverse sequencing and other approaches to compensate for the incompleteness of available maps. Richard Gibbs concurred with the general consensus that sequence generated should be of high quality aiming for 99.99 % accuracy.

Los Alamos National Laboratory

Bob Moyzis stated that the LANL had funding from the U.S. Department of Energy (DOE) for a pilot project on sample sequencing. 1 Mb had been completed which included regions with single-pass sequencing and regions with 16-fold redundancy where they related to EST sequences.

The German Human Genome Programme

Hans Lehrach explained the proposed structure of the German Human Genome Programme which would comprise a central resource centre with genomic and cDNA libraries. The sequencing programme would focus on chromosome 21 and the long-arm of the X chromosome, for which the map was 70% complete.

EMBL

Wilhelm Ansorge provided brief details of the sequencing technology used at EMBL; with 2 dyes and 2 lasers, 2000 bases could be read from each sequencing reaction. He estimated that with this technology 3 people could sequence 4Mb per year at 5-fold redundancy. The Laboratory had focused on the EC yeast genome programme to date and did not have an in-house human mapping programme. Any human sequencing programme would therefore be dependent on resources being made available from other laboratories.

Marshfield Medical Research Foundation

Jim Weber described his whole-genome sequencing strategy which he argued would be more efficient than a clone-by-clone approach. He proposed a scheme whereby the majority of sequence information would be derived by the end-sequencing of clones with an insert size of greater than, or equal to, 5 Kb. He proposed that quality values should be assigned to each nucleotide and that sequence should be deposited, as generated, into a common public repository. Sequence assembly could be carried out by one, or several, large centres. Jim Weber estimated that over 30 Gb of human sequence data will have been generated by 2000 (10-fold coverage). He argued that the whole genome approach would provide a more complete and less artifactual coverage of the genome than the clone-by-clone approach. It would also be achieved at lower cost (0.5 cents per base raw data and 5 cents per base finished sequence).

DISCUSSION

The key issue discussed was the balance between high throughput and resolving difficult genomic regions. Long repetitive elements (LINES) and clone rearrangements were likely to be very difficult to resolve with end-sequencing strategies. Anchoring contigs *via* STSs would assist to some degree but it was agreed that all methods would have some drawbacks and that, at the current time, it was important to pursue multiple approaches. Some participants argued that, in order to achieve the goal of sequencing the human genome by 2001, it would be necessary to focus on developing high throughput rather than resolving difficult regions. Others argued that integrating the sequencing with mapping strategies would resolve many of the problems and would represent only a small fraction of the final cost. In reality, it was recognised that these issues could only be properly addressed when large amounts of human sequence data had been generated by the various approaches and the problems of fidelity had been evaluated.

SESSION II - SEQUENCE-READY MAPS

CHAIR: David Cox

David Cox identified three key issues for speakers to address in relation to the implementation of strategies for large-scale human sequencing. These were:

- The role of STS anchor maps.
- Strategies to develop sequence-ready maps/minimal tiling paths.
- Strategies to ensure that sequence or maps derived from cloned material was actually representative of the genome.

The session was divided into two parts; the first part focused on strategies for large-scale human sequencing based on experience with model organisms, and the second part focused on resources available for human sequencing and how they were being used.

Part I: Model Organisms

John Sulston - *C.elegans*

John Sulston summarised progress and lessons learned from the *C.elegans* sequencing project. The physical map covered more than 95 Mb with 7-deep coverage over 80% of the genome and with 7 gaps. Cosmid contigs had been assembled by fingerprinting and the deep coverage had meant that the tiling path had been easy to determine. YACs had been incorporated by hybridisation strategies but with lower confidence levels because of repetitive sequences. To date, 34 Mb had been sequenced by the two centres (WashU and the Sanger Centre), 6000 genes had been identified of which 45% showed database matches, 28% of the sequence represented coding information with an average gene density of 1 per 5Kb (autosomal) and 1 per 7 Kb (X chromosome), 30% of predicted gene sequences could be confirmed with EST/cDNA matches.

The key features of the strategy were the use of multilevel maps (cosmids and YACs) to resolve difficulties and provide a range of resources for use by the scientific community. The high resolution map had resulted in greater sequencing efficiency and, as a general rule, it was more efficient to aim for a high quality product than attempt to resolve mistakes at a later date.

In response to questions about testing the integrity of the genomic sequence, John Sulston stated that, at one point, cosmids had been tested against the genome by PCR. However, this had been very expensive and was not 100% efficient, it was not therefore considered to be worthwhile. Fingerprinting allowed validation to 3 Kb resolution but he conceded that the cosmid sequence had not been truly validated against the genome. There was some discussion about the mutation rate in cosmids and M13 clones but it was generally agreed that these were fairly simple to detect, particularly in the worm since it was homozygous.

David Cox noted that there were both technical problems and efficiency costs inherent in the validation of sequence information and producing a truly complete product.
John Sulston commented that completion of the human genome would be asymptotic and that some features (such as centromeres) may be omitted by design.

Craig Venter - *H. influenzae*

The whole genome shotgun approach used for *H.influenzae* was based on the use of assembly software to generate a 'sequence map'. The ease with which sequences could be assembled depended on the quality of the sequence data.

Following the initial assembly, gaps were filled either by PCR or by using lambda clones, particularly where they related to physical regions which could not be cloned in *E.coli*. David Cox asked whether the range in size of genome assemblies followed a random distribution; this was affirmed.

The genome sequence of micro-organisms was amenable to validation by restriction digest or PCR, although in some cases it may be difficult to isolate the genomic DNA for this validation. The coding sequence could be edited using a frameshift editor (although some frameshifts may serve a regulatory function and be intrinsic to the sequence).

TIGR was proposing to sequence chromosome 16p of human (30Mb) in collaboration with Bob Moyzis.

The role of long PCR to bridge physical gaps and orient assemblies was discussed. The value of this technology depended on the number of ends/permutations that were involved and it was pointed out that regions of DNA that were difficult to clone in bacteria were also often difficult to PCR.

Craig Venter quoted a current cost at TIGR of 30cents/base part direct costs with 50cents/bp including building costs. He agreed with the general consensus that it was important to follow multiple approaches and to optimize costs.

Part 2: Existing Resources

This session focused on the availability of resources (maps, clones, software etc) and their use in development of large-scale sequencing strategies. All participants were asked to submit a list of resources that they had made available in the public domain. These are attached as an appendix to this report (Appendix 1).

Rick Myers

Rick Myers described the approach at the Stanford Human Genome Centre to generate high resolution radiation-hybrid maps and BAC contigs. A protocol to verify map and sequence data with oligo chips was being developed in collaboration with Affymetrix. The 10bp chips were likely to have a useful role in checking fingerprinting but could only verify sequence data to a limited extent. The chip technology was likely to be prohibitively expensive for 100% verification. In addition, there were serious technical problems resulting from oligo folding that meant that some sequences would not be accessible using this technology. The Stanford Human Genome Centre had initially proposed a strategy based largely on a directed sequencing approach using transposon-mediated directed sequencing but it was now likely that their strategy would also involve a random approach.

David Bentley

David Bentley summarized progress on the development of a sequence-ready map for chromosome 22. The key elements of the approach were the use of a radiation hybrid map to pick out large insert clones and then to integrate PACs, BACs, cosmids etc. *via* fingerprinting.

Eric Lander

The Whitehead Institute/MIT Center for Genomic Research had mapped 16,500 STSs to generate a human physical map with 8,000 mapped to Radiation Hybrid addresses and 11,500 to YACs. This included 60% Genethon and CHLC loci. The Center planned to map a total of 20,000 STSs by June 96.

In the mouse, 6,500 STSs had been mapped to generate a genetic map published in Nature, 14th March 1996. Eric Lander suggested that in order to interpret the human genome effectively, it would be worthwhile sequencing 5% of the mouse genome for comparative studies.

The rapid STS mapping technology could be used to pull out YACs, BACs and PACs for the generation of sequence-ready maps and closure of gaps. The sequencing strategy proposed by Eric Lander was to pull out BACs using STSs to generate end-sequence and fingerprint the BACs to check integrity.

In discussion about fidelity checking, it was agreed that deep maps of at least 10-fold coverage would be required to distinguish between polymorphisms and clone rearrangements. This was particularly true of regions where genomic rearrangements were to be expected.

Pieter de Jong

Pieter de Jong spoke about the PAC and BAC libraries that he had generated. The original PAC library with insert sizes of 110-120 Kb was derived from male DNA and the X-Chromosome sequences were therefore under-represented. The library had been distributed widely and was used by the Sanger Centre for the sequencing of the BRAC2 region. For the future, Pieter de Jong hoped to generate libraries of 10-20 fold redundancy in BAC vectors. Sequencing from PACs presently required a gel purification step to separate insert fragments from vector sequences. However, other technologies were being developed to carry out this separation more effectively.

Mel Simon

Mel Simon pointed out that whilst the insert size of PACs and BACs were similar at 110-150 Kb, the PAC vectors were significantly larger at 11-18 Kb, compared to 7 Kb for BACs.

Mel Simon had used 400 chr 22 markers from the Sanger Centre to generate a BAC map of chr 22 with 625 clones in 120 contigs with an average size of 350 Kb. The map included 556 landmarks of ESTs, STSs and cDNAs. Mel Simon was willing to make the clones available to Bruce Roe and the Sanger Centre for sequencing, if required. Similar collaborations had been established with Lisa Stubbs and Rick Woychik at Oakridge National Laboratory to generate BAC maps and with Craig Venter and Bob Moyzis on chromosome 16 to convert existing map information into BAC contigs.

Glen Evans

Glen Evans described a pilot project to generate a cosmid-based sequence-ready map for chromosome 11. The project involved the end-sequencing of 18,000 cosmids (5-fold coverage). To date, Glen Evans' group had generated 15,000 end-sequences equivalent to 5.9Mb and sequenced 11 cosmids and 1 PAC clone to 99.9% accuracy. The sequences were currently available on the Center's Web Page. Analysis of the data generated so far had shown 24% STSs, 24% ESTs, 15% repeats, 4% known genes, 31% motif-similarities and 16% of known chromosome 11 sequences detected (i.e. 29 of the 174 known genes on chromosome 11 had been detected). Two cosmids had been sequenced by cosmid-oriented walking (cow) but this was a very expensive approach due to the cost of oligos. The Center had therefore developed an automated oligonucleotide synthesizer to generate oligonucleotides directly from the assembly data (PHRAP) at a cost of 5 cents per base. The potential reduction in costs could allow a more directed approach to be used.

In discussion it was recognised that there was general requirement for technological developments such as oligonucleotide synthesizers to be more readily available and distributed. Whilst there was clearly commercial interest in licensing the technology to sell oligonucleotides, this had not facilitated the distribution of the technology itself, which may require the involvement of contract engineering firms.

SESSION III - LARGE-SCALE SEQUENCING

CHAIR - Tom Caskey

The session began with a discussion on data release and led by John Sulston and Bob Waterston. Drs Sulston and Waterston proposed that sequence data should be released automatically on a daily basis and that there should be no patent protection before release. A number of key issues were raised during the discussion:

Data Release

There was a strong scientific argument for immediate data release in order to facilitate co-ordination and encourage further research and development.

The value of unfinished sequence information was queried by some participants but previous experience (e.g.with the BRCA2 region) had shown that such information could be effectively utilised by both academic and commercial groups.

There was a balance between providing high quality information and avoiding long delays in the release of data. It was agreed that early release of data would be essential to ensure that sequence information was freely available for research and development.

IPR

Primary sequence information of unknown function was unlikely to be patentable.

In the U.S. patents could be filed up to one year after data release whereas this was not the case in Europe. This meant that data release would have a different effect on the ability to patent in different countries.

It was important to ensure that centres funded to generate sequence information in the public domain, on a large scale, did not also establish a privileged position in the control and exploitation of that information.

It was noted that groups in different countries and funded by different agencies may be under various legal and political constraints which would make it difficult for them to adopt these principles. It was agreed that funding agencies should be urged to foster these principles in the public interest.

Lee Hood

Lee Hood spoke about the use of the software programmes, PHRED and PHRAP, developed by Phil Green at Seattle for base calling, quality assessment and assembly. The PHRED programme was developed to identify the highest quality data and in conjunction with PHRAP gave an overall discrepancy rate of one base in 2-3 kb before editing. The large-scale sequencing strategy adopted at Seattle involved random shotgun sequencing of BACs at 6-fold to 8-fold coverage with closure using combinatorial PCR. Of the 4.5 Mb sequenced to date, 35% had represented repetitive sequences; software had now been developed to mask and remove 99% of these repetitive sequences.

Rick Wilson

The Genome Sequencing Centre at WashU also used the PHRED and PHRAP programmes to improve the processing and assembly of sequence data from the shotgun stage. A new programme called GETLANES was used for lane-tracking and data-extraction, in preference to the ABI software. PHRED was used for base-calling and assignment of quality values. Another new programme called FINISH is used to automate much of the subsequenct directed sequencing.

Current costs were approximately 40 cents per base, including technology developments, informatics, mapping (at WashU) and equipment amortised over 5 years. The aim was to reduce costs to 10 cents per base and to produce 25 Mb of finished human sequence per year. Recent improvements had included more lanes per gel, more bases per read, higher signal strengths, more automation, more sequence per person and per unit cost.

Rick Wilson was asked about the value of reverse-reads and stated that these tended to be used for finishing in a targeted fashion after PHRAP assembly. For plasmid sequencing it may be useful to include reverse reads as part of the main strategy but, in general, there was a balance between the depth of coverage and the role of reverse reads.

John Sulston

The Sanger Centre had submitted 2.7 Mb of finished human sequence to the public databases. This included 1.8 Mb from the Huntington Disease region which contained at least 20 different genes. The Centre planned a systematic sequencing approach for large regions of chromosome 22, X, 6, 20 and 1. In the next 3 years the Centre planned to increase its output of finished sequence from 20 Mb per year to 100 Mb per year.

Richard Gibbs

At Baylor, Richard Gibbs' group was using new dyes which were not covered by existing patents to provide ET-primers. Reverse reads were used as a part of the overall strategy to reduce the need for oligonucleotide synthesis and to verify overlaps. A region of chromosome 12 near CD4 had been sequenced and analysed. All except one of the genes identified using software tools had been verified by EST hits. The Group was also involved in comparative sequencing of regions of Xq28 in man and mouse; high quality data was a prerequisite for reliable comparison of coding regions.

Trevor Hawkins

Trevor Hawkins described the Sequatron; a fully automated front-end system for the preparation of sequencing reactions. The system was currently capable of processing 8,000 samples per day with a projected increase to 12,000 samples per day. The Whitehead/MIT Center used a base-calling software called GRACE as a substitute for the ABI software. To date, all the sequencing at the Center had been ESTs and STSs from man and mouse but the aim was to produce 5 Mb of finished human sequence in the first year rising to 80 Mb in the third year.

Andre Rosenthal

Andre Rosenthal was the co-ordinator for the large-scale human sequencing programme of the German Human Genome Project. At Jena, he had a group of 37 people of whom 25 were involved directly in sequencing. His laboratory currently had 12 ABI 377 sequencers and he hoped to increase this in 1996/97. He hoped to produce 5-10Mb finished sequence by the end of 1997, 10-15 Mb in 1998 and 15-20 Mb in 1999, assuming that the restrictive regulations on consumable expenditure could be overcome. Funding had been provided by various agencies including the EU, the DFG, the BMBF and the German Human Genome Programme. The German BMBF provided most of his funding. Targeted regions included: 3 Mb of Xq28 between the MeCP2 locus and DXS304 (in collaboration with Annemarie Poustka, Michele D'Urso, the Sanger Centre, Richard Gibbs and Ellson Chen), 3 Mb of Xp11.23, 1-2Mb of Xp11.4, 1-2 Mb in PAR 1 of X, 1-2 Mb regions on chromosomes 7, and 11 around disease genes and fragile sites. Andre Rosenthal was also involved in a project proposal to the German Human Genome Programme to sequence 30-40 Mb of chromosome 21. This proposal excluded the Minimal Downs Syndrome Region and the PME region which were being sequenced by two Japa: ese groups and the Stanford Genome Centre. Rosenthal's group was also involved in comparative sequencing studies in human and fugu for genes on the X-chromosome. In order to investigat gene evolution between man and fugu as well as possible synteny between the two species, he hoped to carry out comparative analysis on larger regions of the human X chromosome in fugu and man. His group also hoped to analyse genes from chromosome 21 in fugu although funding for comparative analysis on a large scale was not available in Germany. He was supportive, in principle, of the data release policy proposed earlier but would prefer high quality data to be released. His group would make all sequence data funded by the German BMBF and the German Human Genome Programme available immediately, preferably on a weekly basis. All the sequence data would have accession numbers and would be submitted simultaneously to GenBank and EBI as well as being accessible *via* the IMB Web page. It was unclear whether the BMBF would impose specific regulations on data release for its grant holders. Rosenthal's personal opinion was that funding agencies should be urged to adopt open data release policies; he hoped, therefore, that the BMBF would not place any restrictions on data release. Salary costs for technicians and post-docs were approximately twice as high in Germany in comparison to the UK and the USA, the lowest possible sequencing cost was, therefore, likely to be of the order of 30 cents per base for the period 1997-1999. In Germany, sequencing costs were lowest at the Institute of Molecular Biotechnology at Jena.

Bruce Roe

Bruce Roe was focusing on chromosome 22 using Gene Myers' assembly programme, FAK II. He had received BACs from Mel Simon nucleating from biologically important regions on chromosome 22. Human-mouse comparative studies were being progressed for the DiGeorge region and other syntenic regions on chromosome 10. 3 Mb had been sequenced to date of which 10% represented Alu sequences. 250 Kb had been submitted to Genbank and a further 1 Mb had been granted accession numbers and would be available within the next fortnight.

Jim Weber

Jim Weber had proposed a whole genome shotgun approach at a cost of $26 million over 4 years; this would provide 4.85 x 10^9 bases of raw output. The direct costs represented 65% of the total and one of the key costs was that of the enzyme. Jim Weber argued that it was important to reduce enzyme costs by developing new ones or by negotiating more strongly with the companies involved. The Marshfield Research Foundation was developing new machines for sequencing and also for PCR reactions.

Masahira Hattori

Masahira Hattori described his nested deletion approach to reduce the number of sequencing reactions required to cover a particular region. His group had produced P1 contigs covering 1.6 Mb of the Down Syndrome critical region, 21q22.2, and planned to sequence the entire region at 99.75% accuracy. Two P1 clones had been sequenced to date and the raw data had been sent to the STA for validation.

SESSION IV - INFORMATICS

CHAIR: David Lipman

In view of the strong consensus on data release achieved the previous day, David Lipman proposed that the session should focus on annotation of data, mechanisms for assessing error rates and new approaches.

Mark Adams

Mark Adams described software developments at TIGR for sample tracking (TRACKER), quality control feedback and assessment of randomness of a particular library.

He also described various ways of annotating sequence data based on similarity searches (e.g. EST hits and BLAST analysis) or gene prediction analysis using GRAIL and other software. Different methods may give different results and it was important that these were annotated appropriately, and any contentions marked, since the average user was unlikely to have sophisticated sequence analysis software or hardware.

It was agreed that detailed annotation submitted to public databases should be definitive and that submitting groups should be responsible for revising any inaccuracies in the sequence or annotation. The NCBI could revise EST hits since it carried out daily comparisons of existing data with new submissions. One of the key issues was the high level of redundancy which was accumulating in the public databases from entries with an increase in the level of experimental and computational information attached which were being considered as independent entries rather than revisions.

LaDeana Hillier

The goal at Washington University Genome Sequencing Center was to provide immediate data release with local annotation of sequence. The units of release were BACs, PACs and cosmids which could be updated into larger contigs.

All features annotated were at a high confidence level; local analysis and annotation were important for assessment of error rates. COP and P-COP were used to automatically compare the consensus sequence against all available raw data to check the consistency of the data. Assemblies were confirmed using mapping data from STSs, restriction digests, fingerprints and overlapping clones. The public availability of raw data also provided an independent checking mechanism..

Polymorphisms in the human genome sequence were annotated in a feature table. These could be distinguished from clone mutations with a high confidence level in the human genome sequence because of the depth of coverage. They were not annotated in ESTs because the error rate of single-pass sequencing was too high for polymorphisms to be distinguished with high confidence.

Chris Fields suggested that sequence data should be annotated with an identifier for the clonal source so that these could be retrieved. It was common practice to archive cosmids and BACS but to dispose of derivative subclones after sequencing. The private sector, particularly Research Genetics, made a valuable contribution to ensuring that source material was widely available. Hans Lehrach suggested that efforts should be made to ensure that different types of experimental data relating to specific clones could be retrieved more easily. The use of universal identifiers, co-ordinated by HUGO, would greatly assist this process.

Jim Weber

Jim Weber's key interest was in the detection of genetic polymorphisms by human genomic sequencing. Polymorphisms present at a level of 10-33% should be detectable with 10-fold coverage.

He proposed a whole genome shotgun approach and described computer simulations for assembly of sequence data derived using this strategy. Simulations had been carried out by Eugene Myers, with the following assumptions: 10-fold coverage, minimum of 35 nucleotide overlaps between sequences and 25% repeats. With 10-fold coverage and sequence from both strands, 17,000 small gaps would be expected. With sequence from only one strand, this would increase to 616,000.

A number of concerns were raised about the feasibilty of this approach in practice. The non-random distribution of repeats in the human genome and the huge scale of the problem meant that the simulations were unlikely to hold true, in practice. In addition, the approach did not include a tested strategy for gap closure which was likely to represent a difficult and expensive part of the approach.

Jim Weber argued that the value of the whole genome shotgun approach was that it would generate large amounts of data very quickly and in a uniform way. There was support for the objective of identifying polymorphisms but this should not be confused with the overall sequencing project. A more cost-effective approach may be to generate single-pass sequence from other individuals for comparison with the high quality reference sequence.

Mark Adams commented that experience from sequencing small geno nes had shown that whilst some genomes followed the Lander-Waterman predictions well, others did not. It was often difficult to assess this until a large amount of data had been collected, it was then very expensive to adopt alternative strategies to resolve these difficulties.

Richard Durbin

Richard Durbin explained how manual operations in the sequencing process were becoming computerised at the Sanger Centre and the need for human input being reduced. The computerisation involved the introduction of new software as modules rather than revision of a single monolithic structure.

He discussed various ways in which confidence levels and error rates could be attributed to data. It was general practice to assign confidence levels for base-calling on a log scale, via PHRAP, rather than use a binary system. Most uncertainties had very low PHRAP scores and tended to be clustered. In order for external users to make best use of the data available, it was important that centres explained how data had been validated and to ensure that PHRAP values (or other confidence levels) and raw traces were also accessible.

SESSION V - PANEL/OPEN DISCUSSION

CHAIRS: Jim Watson, Bob Waterston and John Sulston

The following key issues were identified to be addressed in the summary session:

- Data release and intellectual property rights
- Co-ordination
- Funding available for large-scale human genome sequencing
- Accuracy; standards and evaluation
- Data release and intellectual property rights

John Sulston and Bob Waterston proposed the following principles for release of human genomic sequence generated by large scale centres:

Release

- Automatic release of sequence assemblies greater than 1 Kb (preferably daily).
- Immediate submission of finished annotated sequence
- Aim to have all sequence freely available and in the public domain for both research and development, in order to maximise its benefit to society.

Policy

- The funding agencies are urged to foster these policies.

All participants voted, on a personal basis to endorse these principles. It was noted that some centres may find it difficult to implement thse principles because of legal constraints and it was agreed that all participants should be given an opportunity to comment on the final statement before it was released. The summary statement at the beginning of the report has taken account of comments made since the meeting.

Co-ordination

All participants were invited to state their initial sequencing targets over the next few years. It was recognised that, in most cases, these should be regarded only as intentions and were dependent on funding being made available. These intentions are detailed in Appendix 2 of this report.
In order to promote co-ordination of activities, it was agreed that large-scale sequencing centres should inform HUGO of their intention to sequence particular regions of the genome. HUGO would present this information on their World Wide Web page (http://hugo.gdb.org/hsmindex.htm) and direct users to the Web pages of individual centres for more detailed information regarding the current status of sequencing in specific regions. This would allow centres to declare their intentions in a general framework whilst also allowing more detailed interrogation at the local level.

It was agreed that the aim of encouraging centres to declare interests in particular genomic regions was to increase efficiency and avoid duplication. It was also important that outside users should be kept informed about the status of sequencing of particular regions of the genome. Participants agreed that it was important to develop mechanisms to facilitate these aims.

Funding

Jim Watson posed the question as to whether there was likely to be sufficent funds available to complete the human genome sequence within the next five years.

NIH-NCHGR

Francis Collins stated that the NCHGR currently provided $30 million per annum for sequencing and technology development which would increase to $50 million with the new awards, under the RFA for large-scale sequencing, which would be announced in April. $30 million of this could be attributed to sequence production with the rest going towards technology development. Most of the current funding for sequencing was directed towards model organisms such as *C.elegans* and *Drosophila*. The additional funding under the RFA would allow at least $15 million to be provided for human sequencing this year. This was expected to increase to $60-$80 million (of a total NCHGR budget of $120 million) in 1999 with the move from mapping to sequencing. Grants awarded under the RFA would be reviewed in December 1997 to decide whether a third year of funding should be provided. The assessment criteria for renewal would include the amount of finished sequence submitted to public databases. Early release of data and a high level of accuracy would also be required.

The NCHGR would be holding a workshop of grantees in April to consider potential mechanisms to determine accuracy and validate data. The NCHGR was also planning a symposium on post-genomic biology which would consider: Future technologies in sequencing and genotyping, whole genome approaches to function, evolutionary biology, human variation and the ethical, legal and social issues (ELSI). The mouse genome and yeast whole genome biology were considered to be likely targets for investment in the near future.

The Wellcome Trust

Michael Morgan summarised the Trust's main activities in genome research; the development of the Wellcome Trust Genome Campus at Hinxton, pilot studies on pathogenic genome sequencing and the Wellcome Trust Centre for Human Genetics, in Oxford. The Trust was planning a scientific frontiers meeting "From Gene to Structure and Function" to address potential strategies for genome interpretation and exploitation. The Trust funding for the Sanger Centre totalled £84 million over the next 7 years of which £60 million was attributable to human sequencing and associated mapping; equivalent to £8.5 million per annum.

The Sanger Centre had orginally proposed to sequence a third of the genome and had been funded, by the Trust, to sequence a sixth. John Sulston considered that the Centre could realistically scale up to do 50% of the genome if funding were available.

European Commission

Manuel Hallem stated that the Fourth Framework programme (1994-1998) provided funding of $17,000 million over 5 years of which $11 million per annum was used to support human genome research. This covered human mapping and sequencing, function, disease determinants, gene therapy and data management. An ad hoc working group was currently considering priorities for the Fifth Framework programme. A key criteria for inclusion was that topics should be complementary to activities supported in member states. The first contracts under the fifth framework would be isued from 1st July 1999.

France

Jean Weissenbach stated that France was currently considering the development of a French genome sequencing programme but nothing had yet been agreed.

Germany

Frank Laplace (Federal Ministry of Research and Technology; BMBF) informed participants that a Scientific Advisory Board for the German Genome Programme would convene shortly to initiate the programme. The BMBF would be providing funding of DM 40m-50m per annum which would include support for two resource centres to be directed by Hans Lehrach and AnneMarie Poustka. The Deutsches Forschung Gemeinschaft would be providing an additional DM 5m-10m for genome studies focussed on the identification of disease genes. It was hoped that additional funds would be provided *via* investment from industrial partners and discussions were currently in progress with this aim. Industrial participants were requesting privileged access to data for three months prior to publication but this was currently the subject of further negotiations. Notwithstanding industrial sponsorship, Frank Laplace endorsed the principle that work funded with public money should be in the public domain.

U.S. Department of Energy

David Smith stated that the DoE budget for the human genome programme in 1996 was $70m per annum of which $10m was attributable to human and mouse sequencing and $15m to development of new sequencing technologies. In addition to this funding, the DoE also provided $4m per annum in support of microbial genome sequencing.

U.K. Medical Research Council

Sohaila Rastan stated that the MRC currently provided support for the C.elegans sequencing programme at the Sanger Centre at the level of £13.1m over 5 years (1993-1998). In addition, a further £10m would be available over 5 years from 1995 for genome research at the Sanger Centre; £2m of which would be used to ramp up and complete the C.elegans genome sequencing project. The remainder would go towards the human sequencing programme.

Accuracy

It was agreed that sequencing centres should aim to achieve 99.99% accuracy.

Discussion focussed on measures that might be required to achieve this level of accuracy and the cost/benefit ratio of the various methods. These included:

- Double-stranded coverage
- "Rule of Three": i.e. two clones including one reverse-read or using orthologous chemistry
- Resolution of all ambiguities
- High level of contiguity

It was noted that some regions may require additional reads to achieve this level of accuracy and others possibly less. The quality of the data could be determined by the ease of assembly and the use of software programmes such as cop and pcop which compared the consensus sequence with the raw data. Other methods of quality control which were discussed include the resequencing of a proportion of clones, independent analysis of trace data, and comparison of assembly data with restriction analysis. It was noted that data quality was likely to vary depending on the base composition of particular regions of the genome. Sampling would therefore have to be quite extensive in order to provide a comprehensive picture.

In considering the level of contiguity that might be achieved, it was noted that sequence "gaps" arose for three main reasons; "biological" cloning gaps, technical gaps arising from dinucleotide repeats or G,C-rich regions, and sizing or mapping gaps. In some instances, it may be necessary to develop further technologies to deal with the problems and it was therefore agred that gaps should only accepted if all exisiting technologies had been exhausted.

Participants were informed that the NIH NCHGR would be convening a workshop of grantees to discuss validation and quality control of data in April.

SANGER CENTRE (22.2.95)

PROGRESS IN FLOW-SORTED CHROMOSOME LIBRARY CONSTRUCTION

Chromosome	Sorting	Sort purity	Theoretical Coverage	Library purity (FISH)	Chromosome equivalents picked
○ **chromosome 22**					
cosmid library	completed	98%	23x	40/40	10
○ **chromosome X**					
cosmid library	completed	96%	23.5x	30/30	in progress
fosmid library	underway	96%	3x		
○ **chromosome 6**					
cosmid library 1	completed	97%	45x	39/40	
cosmid library 2	completed	97%	7.5x		
fosmid library	completed	97%	9x	5/5	
○ **chromosome 7**					
cosmid library	completed	97%	15x	39/40	
fosmid library	underway				
○ **chromosome 5**					
cosmid library	completed				
fosmid library	underway				
○ **chromosome 20**					
cosmid library	completed				
fosmid library	underway				

RESOURCES ·

Sanger Centre / D. Bentley / R. Durbin / J. Rogers / J. Sulston

Mapping reagents, data, software:

New clone libraries : flow-sorted chromosome cosmid + fosmid
libraries - see back of sheet.
Freely available, subject to time to
pick + copy.

Data: Chromosome 22 map acedb (22ace) on ftp site
in databases Integrated multilevel map markers + YACs.
Clones etc available : contact idl@sanger.ac.uk.

X chromosome map information Xace
available soon. : contact mtr or drb@sanger..

PACE. Clone supply status, minimum tiling
paths, integrated with long-range
landmark map. 22 + X regions.
on WWW. Contact sd, rd@sanger

RH maps data being put into EBI
Contact panos@sanger

Sequence data on ftp and submitted : see separate sheet
and also contact rd@sanger . . .

Contigs >1kb on ftp overnight: jkb@sanger . .

Worm db. acedb and other

[illegible]

RESOURCES AVAILABLE

Name of participant : Mark Adams / Craig Venter - TIGR

Nature of resources available (software, maps, clones etc.)

Software

TIGR Assembler - sequence assembly

HBQCM - hexamer-based composition tool

yank - GenBank extraction software

TIGR's sybase schema

Human cDNA Database - >355,000 ESTs, >51,000 THC assemblies

Available via:

e-mail request to arkerlav@tigr.org (Tony Kerlavage)
cDNA clones through TIGR/ATCC and www

Any conditions attached:

None.

RESOURCES AVAILABLE

Name of participant : ANSORGE/EMBL

Nature of resources available (software, maps, clones etc.)

- SEQUENCING TECHNOLOGY – 100kb/per run

GENESKIPPER – ASSEMBLY PROGRAMM
+ SEQUENCE ANALYSIS

RAN-DI (Random-Direct) strategy –
- assembling first 80-100 clones randomly sequenced
+ all ECOR1 fragments, finish with DIRECT strategy
→ NO CLONING GAPS OBSERVED.

Available via: FAX or by mail to EMBL

Any conditions attached:

RESOURCES AVAILABLE

Name of participant :

Tony Carrano
Lawrence Livermore National Laboratory

Nature of resources available (software, maps, clones etc.)

Resource	Availability
High-resolution, metric map of chromosome 19	Published version available in Dec issue of Nature Genetics. Detailed version available by collaboration
Arrayed cosmid libraries of human chromosomes	Through major genome centers. Soon to be available through the UK and German resource centers.
IMAGE collection of cDNAs	Available through industry and resource centers.
DNA sequence sample tracking software	Contact Tom Slezak @ LLNL
Clone fingerprinting assembly and database software	Contact Tom Slezak @ LLNL
Mapping infrastructure resource (creating high-resolution sequence ready maps in cosmids and BACs)	Contact Tony Carrano @ LLNL

Available via:

see above

Any conditions attached:

Creating maps as part of the mapping infrastructure resource would require funding.

RESOURCES AVAILABLE

Name of participant : RICHARD DURBIN

Nature of resources available (software, maps, clones etc.)

SOFTWARE; ACEDB database system

SAM (Cari Soderlund) marker assembly/edit/viewer

FPC (" ") fingerprint " " " <u>SOON</u>

AUTOEDIT – sequence automatic editor for assemblies (Richard Mott)

MSPCRUNCH/BELVU/DOTTER – sequence analysis/viewing tools

Available <u>via</u>: ANONYMOUS FTP (FTP.SANGER.AC.UK)

email: RD@SANGER.AC.UK

Any conditions attached:

NO COMMERCIALISATION (use by companies OK)

Resources Available

Name of participant:

Glen A. Evans

Nature of Resources:

1. Chromosome 11 Sequencing DataBases

   YAC/STS coordinates database
   cosmid end sequence database
   YAC-cosmid coordinate database
   Primers (new STSs)
   Homology/Identities listed by match significance

   Chromosome 11 sequencing data (complete cosmid/PAC sequences)
   11p15 project, 11p12 project
   WWW http://mcdermott.swmed.edu/
   Genbank

2. Clone libraries

   chromosome 11 cosmid 5X, arrayed
   chromosome 11 YAC ?X, arrayed (T. Shows/N. Nowak, RP)
   chromosome 11 and 15 PAC set in preparation

   (can be made available on request to G. Evans)

3. Software

   Mermade driver software for 192 channel synthesizer
   Primer prediction software for primer directed walking
   SUMU Lab sample tracking software
   Robotics control software for Biomek
   Data Inspector software for sequence quality control

   WWW http://mcdermott.swmed.edu/

4. Hardware specifications and construction plans

   Prepper III miniprep robot
   Mermade 192 channel oligonucleotidesynthesizer
   Lab workstations
   TREC multigel controller

Lab workstation plans and ordering information

WWW http://mcdermott.swmed.edu/

Available via:

WWW http://mcdermott.swmed.edu/

Any conditions attached:

Data resources are made available within 6 months after generation.

Hardware and software are supplied without warranty and without support other than helpful hints when needed. Hardware specifications and plans are available to all non-commerical users.

RESOURCES AVAILABLE

Name of participant : Chris Fields

Nature of resources available (software, maps, clones etc.)

Chris Fields

GSDB (complete, genome-scale relational DB)
scheduled for operational mid-summer

GSDB "Annotator" multiplatform client
interface (view/edit) available free
mid summer.

Available via: http://www.ncgr.org

Any conditions attached: none

RESOURCES AVAILABLE

Name of participant : Richard A. Gibbs

Nature of resources available (software, maps, clones etc.)

- X chromosome mapped reagents - including binned cosmids (≥ 2,000) -
- Sequences, cosmids and the shotgun libraries from >1mb of human DNA from X, ch12 + ch17 available,
- matched cosmid/cDNA pairs available from X-chromosome, from C.C. Lee.

Available via:

All X chromosome + ch12 resources are described in their respective web pages.

Any conditions attached:

NO.

RESOURCES AVAILABLE

Name of participant : Trevor Hawkins

Nature of resources available (software, maps, clones etc.)

~~>15,000 Human mapped STSs~~

>6,500 Mouse mapped SSRs

GRACE/BASS Gel analysis and basecalling software, UNIX based.

Primer Picking software (PRIMER 2.2)

Lab Base database system

Available via: http://www-genome.wi.mit.edu

Any conditions attached: None.

RESOURCES AVAILABLE

Name of participant : LaDeana Hillier

Nature of resources available (software, maps, clones etc.)

SOFTWARE: GETLANES (tracking gel images)

RETRAK (UNIX interface for editing lane tracking)

TPP (trace processing software)

PHRED (base calling)

PHRAP (sequence assembly)

FINISH (following shotgun completion, finish selects reads to contiguate & improve sequence quality)

DACE (implementation of a laboratory notebook tracking system in ACEDB)

other software tools are also available

Available via: HTTP://genome.wustl.edu/gschmpg.html

PHRED & PHRAP available: phg@u.washington.edu

ACEDB code available: ncbi.nlm.nih.gov : /pub/repository/acedb

Any conditions attached:

retrak & tpp are still under intensive development.

RESOURCES AVAILABLE

Name of participant : PIETER DE JONG

Nature of resources available (software, maps, clones etc.)

120 kb average insert

Human PAC library (16-fold redundant,
(male donor, ~1200 384 well dishes)
DNA from blood)

Human PAC library (~5-fold redundant)
not yet arrayed ; 150 kb insert.
(female donor, DNA from blood

Human BAC library : in progress,
expect to deliver 10-fold redundant
by May '96 and 20-fold by Summer '96.

Available via: PdJ, Roswell Park Cancer Institute

Any conditions attached:

- No secondary distribution of library, no problems to distribute individual clones (no ties attached).
- cost-recovery of labor/plasticware/ mailing costs for library replicates.

RESOURCES AVAILABLE

Name of participant: Dr. Hans Lehrach

Nature of resources available (software, maps, clones etc.)

The Resource Centre distributes high-density gridded filters of genomic libraries, cultures of individual library clones, or (in the future) PCR pools.

The table below gives details of those genomic libraries for which this service is now available, in the near future this will be supplemented with libraries from the I.M.A.G.E. consortium:

Library name	Description	Number
Cosmid (Human)		
L4/FS1	Chromosome 1 specific cosmid library	112
L4/FS6	Chromosome 6 specific cosmid library	109
L4/FS7	Chromosome 7 specific cosmid library	113
L4/FS11	Chromosome 11 specific cosmid library	107
L4/FS13	Chromosome 13 specific cosmid library	108
L4/FS17	Chromosome 17 specific cosmid library	105
L4/FS18	Chromosome 18 specific cosmid library	111
L4/FS21	Chromosome 21 specific cosmid library	102
L4/FS22	Chromosome 22 specific cosmid library	106
L4/FSC X/LA	Chromosome X specific cosmid library	101
L4/FSC X	Chromosome X specific cosmid library	104
Cosmid (other)		
L4/S.Pombe	S.pombe specific cosmid library	60
L4/B/S.Pombe	S.pombe specific cosmid library	61
Fugu-Cosmid	Fugu DNA partial cut with MboI in Lawrist4 and DH10B	66
P1		
P1 Human	Total Genomic P1 Human Library	700
MP1 Mouse P1 library	Total Genomic Mouse C57/Black6 P1 Library	703
pomP1	Schizosaccharomyces pombe (wt 972 h-) P1 library	705
PAC		
Human PAC	Human PAC library brought by Peter de Jong	704

RESOURCES AVAILABLE (Continued)

Name of participant: Dr. Hans Lehrach

Nature of resources available (software, maps, clones etc.)

(Continued from previous page)

Library name	Description	Number
YAC (Human)		
4X YAC	Human YAC library	900
4Y YAC	Human YAC library	901
CEPH YAC	Human CEPH YAC library	904
LSXY	Human YAC library	912
C3H YAC	Mouse YAC library	902
YAC (other)		
St.Marys Mouse YAC RAD52	Mouse YAC library from female C57BL/10 in host strain which is recombination deficient due to mutation in RAD52	909
C57 YAC	Mouse YAC library	903
Whitehead Mousc YAC I	Large insert Mouse YAC library constructed at the Whitehead Institute for Biomedical Research/MIT Center for Genome Research	910
pomYAC	Schizosaccharomyces pombe (wt 972 h-) YAC library	913
ICRF Pig YAC	Pig YAC library	907
LMUB Pig YAC	Pig YAC library from Lymphocytes (~300KB average inserts) from Ludwig Maximillian Univ.Muenchen	911
cDNA (Human)		
Human fetal brain cDNA	Human foetal brain cDNA made from 17 week embryo polyA+RNA	507
HFL cDNA	cDNA using dT primed polyA+ purified RNA from 21 weeks old human fetal liver	512
HTE cDNA	cDNA using dT primed polyA+ purified RNA from 21 weeks old human fetal thymus	508
HPO cDNA	cDNA from 21 weeks human foetal lung, poly dT primed, directionally cloned, excise enzyme MluI	515
cDNA (other)		
MBR cDNA	Mouse adult brain cDNA,synth: oligo dT primed,directionally cloned; cloning site: NotI/SalI; 1.5kb average insert size	510

RESOURCES AVAILABLE (continued)

Name of participant: Dr. Hans Lehrach

Nature of resources available (software, maps, clones etc.)

(see previous pages)

Available via:

The Resource Centre/Primary Database of the German Human Genome Project,
Max-Planck-Institut für Molekulare Genetik,
(Abteilung Lehrach),
Ihnestraße 73,
14195 Berlin (Dahlem)
GERMANY

Tel: [redacted]

Fax: [redacted]

WWW: http://rzpd.rz-berlin.mpg.de/

Any conditions attached:

Distribution of these resources will be free of charge to all participants in the German Human Genome Project, otherwise charges will be made to cover manufacturing expenses and postage costs.

In the case of some libraries additional conditions governing usage and distribution have been imposed by the owners.

RESOURCES AVAILABLE

Name of participant : DAVID J. LIPMAN

Nature of resources available (software, maps, clones etc.)

Databases & Software
see: http://www.ncbi.nlm.nih.gov

Available via: WWW, FTP, CDROM

Any conditions attached: none

RESOURCES AVAILABLE

Name of participant:

Dr. Robert K. Moyzis
Center for Human Genome Studies
Los Alamos National Laboratory
Los Alamos, New Mexico 87545

Ph: 505-667-3912
FAX: 505-667-2891
email: moyzis@telomere.lanl.gov

Nature of resources available (software, maps, clones, etc.)

A) Complete digest libraries for each human chromosome
B) Partial digest phage and cosmid libraries for approximately half of the human karyotype (phage: 4, 5, 6, 8, 11, 13, 16, 17, X; cosmid: 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 20, X, Y)
~~C) YAC libraries for human chromosomes 9, 12, 16 and 21~~
D) M13/STS libraries (can be constructed for any human chromosome)
E) High-resolution YAC/STS/cosmid maps of human chromosomes 5 and 16

Available via:

A) American Type Culture Collection
B) Request from Los Alamos. Will also be available from commercial sources
C) Request from Los Alamos
D) Collaboration with Los Alamos
E) htpp://www-ls.lanl.gov; GDB and GSDB; request materials from Los Alamos

Any conditions attached:

A) Small fee; agreement to acknowledge Los Alamos in publications
B) Must sign Material Transfer Agreement with University of California limiting use to scientific purposes, limiting further distribution and agreeing to a limited collaboration with Los Alamos investigators
C) Collaboration with Los Alamos
D) Collaboration with Los Alamos
E) Sequencing coordinated with Los Alamos

RESOURCES AVAILABLE

Name of participant : Richard Myers + David Cox

Nature of resources available (software, maps, clones etc.)

- two panels of whole genome radiation hybrid DNAs
  (Stanford G3 panel - 400 kb resolution)
  (Stanford TNG panel - 100 kb resolution)
  available from Research Genetics
- map positions of 7300 STSs on the G3 radiation hybrids
- an email server allowing anonymous STS radiation hybrid scores to be integrated ~~on the~~ with our mapping data on the G3 hybrids

Available via: http://www-shgc.stanford.edu

Any conditions attached:

- none

RESOURCES AVAILABLE

Name of participant : Bruce Roe

Nature of resources available (software, maps, clones etc.)

Laboratory Protocols

Cosmid, P1 and BAC'sequence data (In progress)

Available via: HTTP://dna1.chem.uoknor.edu

Any conditions attached:

Let us know if you find something cool that we missed

RESOURCES AVAILABLE

Name of participant : Melvin I. Simon

Nature of resources available (software, maps, clones etc.)

1. Mouse 129ES Cell - BAC Library 235,000 clones (~10X coverage)
2. Human Fibroblast - BAC Library B 70,000 clones (~3X coverage)
3. Human Sperm BAC Library C 75,000 clones (~3X coverage)
4. Human Primary Fib. BAC Library A 100,000 clones (~4X coverage)
5. Human Sperm BAC Library D 75,000 clones
6. 619 - Ch 22 Specific Mapped BAC clones

Available via:

1, 2 and 3 Now Available - Research Genetics Inc (Huntsville Ala)

5 & 6 Available - Research Genetics Inc (April 1996)

4 - Available for screening via Hiroaki Shizuya - Biology Division Caltech - PASADENA
FAX - (818) 796-7066
Also see :-
http://www.tree.caltech.edu

Any conditions attached:

No conditions or restrictions are attached to the material.

RESOURCES AVAILABLE

Name of participant :

Jim Weber

Nature of resources available (software, maps, clones etc.)

Crude, but comprehensive human linkage maps

STRP information

Methods

Image analysis software

Construction information for water bath thermal cycler and some SCAFUD components

Sequence assembly simulation program (from Gene Myers at University of Arizona)

Available via:

Website: http://genetics.mfldclin.edu
Email: gene@cs.arizona.edu

Any conditions attached:

Software is not supported.

RESOURCES AVAILABLE

Name of participant :

Jean Weissenbach

Nature of resources available (software, maps, clones etc.)

The Généthon Human Linkage Map
(5,264 microsatellite markers)

Map + description of reagents
(sequences, primers, alleles, frequencies, etc.)

Available via:

http://www.genethon.fr

Any conditions attached:

freely available